Spring 2015

Profiling gene expression during early gametophyte development and sex determination in Ceratopteris richardii

Nadia Atallah
Purdue University

Graduate School Form 30
Updated 1/15/2015

PURDUE UNIVERSITY
GRADUATE SCHOOL
Thesis/Dissertation Acceptance

This is to certify that the thesis/dissertation prepared

By Nadia Atallah

Entitled
PROFILING GENE EXPRESSION DURING EARLY GAMETOPHYTE DEVELOPMENT AND SEX DETERMINATION IN CERATOPTERIS RICHARDII

For the degree of Doctor of Philosophy

Is approved by the final examining committee:

Jo Ann Banks, Chair	Olga Vitek
Joseph Ogas	Milos Tanurdzic
Michael Gribskov	
Peter B. Goldsbrough	

To the best of my knowledge and as understood by the student in the Thesis/Dissertation Agreement, Publication Delay, and Certification Disclaimer (Graduate School Form 32), this thesis/dissertation adheres to the provisions of Purdue University's "Policy of Integrity in Research" and the use of copyright material.

Approved by Major Professor(s): Jo Ann Banks

Approved by: Peter B. Goldsbrough 4/7/2015
Head of the Departmental Graduate Program Date

PROFILING GENE EXPRESSION DURING EARLY GAMETOPHYTE DEVELOPMENT AND SEX DETERMINATION IN *CERATOPTERIS RICHARDII*

A Dissertation

Submitted to the Faculty

of

Purdue University

by

Nadia Atallah

In Partial Fulfillment of the

Requirements for the Degree

of

Doctor of Philosophy

May 2015

Purdue University

West Lafayette, Indiana

For Alex, mom, dad, and Christina

ACKNOWLEDGEMENTS

I would like to express my sincere thanks to my faculty advisory committee, Drs. Jo Ann Banks, Peter Goldsbrough, Michael Gribskov, Joe Ogas, Milos Tanurdzic, and Olga Vitek for their continued support, expertise, and advice throughout my graduate studies. I would like to thank Dr. Banks, my major professor for professional, academic, and personal support; I could not have asked for a more wonderful role model, advisor, and friend. I would like to thank Dr. Michael Gribskov for teaching me to program, for providing custom Perl scripts during the course of my research, and for inviting me to your fun parties. I owe thanks to Dr. Clint Chapple for providing me with training and time on the qRT-PCR instrument in his lab as well as Dr. Vikki Weake for providing me with training and time on the QIAgility instrument. I would like to thank Dr. Guri Johal for his mentorship and guidance as well as Tyson McFall for helping me to navigate through graduate school. I am grateful to the Purdue Genomics Center, particularly Dr. Phillip San Miguel and Dr. Rick Westerman for the sequencing and their expertise. I would like to thank the Purdue Rosen Center for Advanced Computing for giving me access to the amazing computing facilities at Purdue. I would especially like to thank Steve Kelley for providing invaluable technical support to me. I owe many thanks to Mir Asgar, Andy Eller, Kye Stachowski, and Katie Embry for tissue prep, primer design,

cloning, and much, much more. Thank you to Yuchen Gang for being a wonderful lab mate and for adding to the wonderful atmosphere of the lab.

I would like to thank my family for all the patience, love, and support they have given me throughout the years. To Alex, I could not ask for a more intelligent and fun husband. To my parents, Mike and Karen Atallah, I cannot thank you enough for all you have done for me throughout the years. Thank you for all your emotional (and financial) support. To Christina, I could not ask for a better sister. Thank you for all our Skype chats, mixed CDs, and for simply being there through absolutely everything. I would also like to thank Lady, Kalila, and Watson for always be there to cheer me up when I am stressed and for helping ensure that I do not take myself too seriously.

And finally I would like to thank my friends, especially Jessica Johnston, Katie Duket, Andrea Grovak, Katie Pechin, Sandy Stortz, and Patti Deevy. I could not ask for better, more supportive friends. Each one of you has helped me immensely during this journey. Thank you for helping me get through graduate school with my sanity somewhat intact.

TABLE OF CONTENTS

Page

Page

LIST OF TABLES

LIST OF FIGURES

Figure Page

ABSTRACT

Atallah, Nadia Ph.D., Purdue University, May 2015. Profiling Gene Expression During Early Gametophyte Development and Sex Determination in *Ceratopteris Richardii*. Major Professor: Jo Ann Banks.

In the fern *Ceratopteris richardii*, every spore has the potential to develop as either a male or hermaphroditic gametophyte. Gametophyte sex is determined by a GA-like pheromone (A_{CE}) that is secreted by hermaphrodites approximately 6 days after spore inoculation and induces male development in other juvenile gametophytes. Our goal is to better understand the genetic and molecular mechanisms involved in sex determination and to identify sex determination genes in Ceratopteris. RNA-Seq was used to create *de novo* transcriptome assemblies from gametophytes grown, with or without A_{CE}, during the time that their sex is determined, and from male gametophytes in early development. We found that A_{CE} alters the expression of 1,163 genes, including those involved in epigenetic reprogramming of the genome. This suggests that epigenetics plays an important role in the early establishment of the male program of expression. We also found that a large number of transcripts are stored in the dormant spore (18,437) and that the transcriptomes of male gametophytes early in development are incredibly dynamic. The research presented in this thesis was used to generate easily testable hypotheses and to identify candidate sex-determining genes that had been genetically characterized previously. We propose that the *HERMAPHRODITIC* gene encodes GID1, the A_{CE}

receptor, that the *TRANSFORMER* gene encodes a DELLA protein, and that the *FEMINIZATION* (*FEM*) gene encodes a MYB transcription factor. We also propose that *FEM* directly or indirectly blocks A_{CE} synthesis in the male by down-regulating the expression of a gene (*CPS/KS*) that is essential for GA biosynthesis.

CHAPTER 1. SEX DETERMINATION MECHANISMS IN LAND PLANTS

1.1 Introduction

In all sexually reproducing plants, sex determination is a necessary and important part of the life cycle. It is thought that dioecy in plants, (separate male and female individuals) has evolved repeatedly and independently, as dioecy occurs in the majority of plant orders and appears to be an apomorphy within each order (reviewed in (Charlesworth, 2002)). Consistent with this theory, a diverse range of determinants and processes are involved in sex determination in plants, from sex being determined through sex chromosomes in *Silene latifolia* (Blackburn, 1923), by a combination of hormonal regulation, microRNA, and sex determination genes in *Zea mays* (reviewed in (Irish, 1999; Yamasaki et al, 2005)), to sex being determined epigenetically, based on social environment, such as in *Ceratopteris richardii* (reviewed in (Atallah & Banks, 2015; Tanurdzic & Banks, 2004)). As important as sex determination is in plants, much less is known about sex determination in plants than in animals. For example, comparatively little is known about the structure, molecular function, and maintenance of plant sex chromosomes compared to animal sex chromosomes. Likewise, relatively few sex determination genes have been cloned from plants, and little is understood about the molecular mechanisms controlling sex determination in plants. For this reason and due to the diversity of sex determination mechanisms in land plants, to reach a true

understanding of the mechanisms involved in sex determination in plants, sex determination in a variety of species of plants must be studied.

How, when, and where sex is determined varies greatly among plants and, for this reason, sex determination is difficult to define. For the purposes of this chapter, I define sex determination to be a developmental decision that leads to the differentiation of gamete producing structures. While the life cycles of all land plants involve the alternation between the diploid sporophyte generation and the haploid gametophyte generation, plants have two variations on the life cycle – they can be heterosporous or homosporous (Fig.1.1). While sex determination varies greatly between plants that are heterosporous (plants that produce more than one type of spore) and those that are homosporous (plants that produce one type of spore) (Bateman, 1994; Sussex, 1966), sex determination in either system can be thought of as the decision to make gamete-producing structures. In heterosporous plants, such as angiosperms, this decision is made in the sporophyte generation, whereas in homosporous plants, it is made in the gametophyte, with the production of egg and sperm-forming gametangia, archegonia and antheridia, respectively (Fig.1.1).

In this chapter, recent advances and studies aimed at gaining a deeper understanding of sex determination in plants at a genetic and molecular level are reviewed. Due to the wide variety of sex determining mechanisms throughout the plant kingdom, sex determination mechanisms of representatives from several major clades are discussed to provide a comprehensive view of sex determination in plants.

1.2 Sex determination in angiosperms

The majority of angiosperms (72%) grow perfect flowers, which produce both male and female organs. In these plants, I argue that sex determination can be regarded as the process that regulates the formation of the male reproductive structures (and microspores) and the female reproductive structures (and megaspore mother cells), or as the events/processes leading to the development of heterogametes (Bai & Xu, 2012). The remaining angiosperms are either monoecious or dioecious. Monoecious plants develop with both male and female flowers on the same plant (thus flowers are unisexual but the plants are not) and sex determination is spatially patterned. Some examples of monoecious plants are maize (*Zea mays*), cucumber (*Cucumis sativus*), and fig (*Ficus carica*). Dioecious species are those in which unisexual plants produce unisexual flowers, with male and female flowers growing on separate plants (Seiji Yamasaki et al., 2005). White campion (*Silene latifolia*), garden sorrel (*Rumex acetosa*), and mercury (*Mercurialis annua*) are examples of dioecious plants (S. N. Bai & Xu, 2012). In angiosperms, as in the rest of the plant kingdom, a wide variety of sex determination mechanisms exist. Plant hormones have many effects on plant growth and development, and some of these hormones can also have an effect on sex determination in monoecious and dioecious species (Tanurdzic & Banks, 2004). There is no one hormone that controls sex determination in all angiosperms, and, likewise, the same hormone can have very different effects in terms of sex determination in different species of plants. GA (gibberellic acid) promotes the development of female flowers in maize and yet promote the development of male flowers in cucumber. Additionally, in a number of angiosperms, sex chromosomes have been found to be responsible for sex determination.

1.2.1 The monoecious angiosperms

Zea mays (maize) is a monoecious plant in which sex determination has been well studied. In maize, only unisexual flowers are produced, and they develop in separate inflorescences: the terminal tassels are male and the lateral ears are female. In maize, both the ear and the tassel inflorescence are composed of a spikelet with two glumes (bracts) enclosing two florets (primary and secondary florets) (Fig.1.2A). As spikelets mature, each floret produces a lemma, a palea, three stamen initials, and a gynoecium (Bonnet, 1940; Calderon-Urrea & Dellaporta, 1999; Cheng et al., 1983; Yamasaki et al., 2005). It is after this bisexual stage, during which the ear and tassel florets are morphologically indistinguishable, that sex determination occurs. Sex determination in maize occurs through selective abortion based on the location of the florets in the tassel or the ear: flowers develop from floral meristems that are initially perfect, with both stamen and pistil primordial, and in later development the stamens or pistil primordia are aborted, creating unisexual flowers. Thus, in the tassel, the pistil primordia are aborted (Fig.1.2B) and in the ear, the stamen primordia are aborted (Fig.1.2C) (Bonnet, 1940; Calderon-Urrea & Dellaporta, 1999; Cheng et al., 1983; Kellogg & Birchler, 1993; Kim et al., 2007). The process of sex differentiation in maize does not simply involve abortion of stamen/pistil primordia, but also drastic differences in the structure and pigmentation of the influorescences, and even in the vegetative parts of the plant near these influorescences (reviewed in (Irish, 1999; Yamasaki et al., 2005)). Thus the genes involved in sex determination in maize must control the differentiation of vegetative tissues, pigmentation, and the selective abortion of reproductive organs based on the location of the florets in the tassel or the ear.

Sex determining mutants can provide the basis for understanding the genes and the molecular mechanisms involved in sex determination in maize. Two major types of sex determining mutants have been discovered in maize: those that feminize the tassels and those that masculinize the ears. A number of mutants that masculinize ears have been isolated and characterized. The single-gene, non-allelic recessive *dwarf* (*d1*, *d2*, *d3*, and *d5*) mutants and the *anther ear1* (*an1*) mutant masculinize ears by preventing stamen primordia abortion in the ear (Fujioka et al., 1988; Phinney, 1982; Tanurdzic & Banks, 2004). These mutants are GA deficient and all encode enzymes involved in GA biosynthesis (Bensen et al., 1995; Fujioka et al., 1988). The dominant dwarf mutation *D8* has a very similar phenotype, and encodes a protein orthologous to the Arabidopsis *GIBBERELLIN INSENSITIVE* (*GAI*) gene and the wheat *Reduced height-1* (*Rht-1*) genes, which encode members of a family of transcription factors known to negatively regulate GA response in plants (J. Peng et al., 1999). These mutants provide evidence that GA is involved in the abortion of stamen primordia. Another masculinizing mutation is *silkless1* (*sk1*) (D. F. Jones, 1925). The *silkless1* (*sk1*) gene product blocks cell death and is required for the development of the pistil primordia in the primary ear florets (Calderon-Urrea & Dellaporta, 1999; D. F. Jones, 1925). Maize *sk1* mutants have normal tassels, but have ears in which both stamen primordia and pistil primordia have been aborted (Irish, 1999; D. F. Jones, 1925).

Mutants that feminize the normally male tassels, leading to tassels producing functional pistillate florets, have also been discovered and are known as *tasselseed* (*ts*) mutants, of which 6 loci have been identified: the recessive *ts1*, *ts2* (Emerson, Beadle, & Fraser, 1935), and *ts4* (Phipps, 1928), the dominant *Ts3* and *Ts6*, and the semi-dominant

Ts5 (Emerson et al., 1935; Irish, 1999; Nickerson & Dale, 1955; Seiji Yamasaki et al., 2005). The *ts1* and *ts2* mutants display particularly dramatic feminization phenotypes; these genes are required for the death of pistil cells and thus feminize the tassel, converting all tassel florets from staminate to pistillate (Calderon-Urrea & Dellaporta, 1999; Irish, 1999; Nickerson & Dale, 1955; Seiji Yamasaki et al., 2005). Additionally, these mutations lead to development of a double-kerneled spikelet in the ear, due to the successful development of the second floret in the ear spikelets (Calderon-Urrea & Dellaporta, 1999). TS1 is involved in an early step in the biosynthesis of jasmonic acid (JA) and the ability of applied JA to rescue stamen development in *ts1* and *ts2* mutants suggests that both *ts1* and *ts2* may be involved in JA biosynthesis (Acosta et al., 2009). The *tasselseed2* gene encodes a short-chain alcohol dehydrogenase/reductase with broad substrate specificity (DeLong, Calderon-Urrea, & Dellaporta, 1993; Wu et al., 2007). In 2007, Hake *et al.* found that *tasselseed4* is a miR*172* microRNA that targets an *APETALA2*-like floral homeotic transcription factor (Chuck, Meeley, Irish, Sakai, & Hake, 2007). Thus, microRNAs are involved in sex determination and development of the tassel. Recently, another mutant that feminizes tassels and also effects the stature of the plant, has been investigated and has provided evidence that sex determination in maize tassels may be controlled by another class of phytohormone, brassinosteroids (BRs). The *nana plant1* (*na1*) mutant is a dwarf mutat caused by the alteration of a 5α-steroid reductase – an enzyme involved in the biosynthesis of brassinosteroid (Hartwig et al., 2011).

We now know that sex determination in maize is a complicated process that involves the interplay of phytohormones as well as genetic control, and the action of

microRNAs. However, we do not know how these hormones regulate sex, or what genes are involved. Future studies to identify genes that respond specifically to GA to induce stamen primordia abortion would be useful, as well as studies to elucidate the molecular and genetic basis for the effects of BRs on sex differentiation.

Another monoecious plant that has been used extensively for research on sex determination in plants is *Cucumis sativus* L. (cucumber), which belongs to the Cucurbitaceae family. Though most cucumber plants are monoecious, depending on genotype they can be also be hermaphroditic (produce bisexual flowers), gynoecious (produce only female flowers), androecious (produce only male flowers), and andromonoecious (produce a combination of male and bisexual flowers) (Malepszy & Niemirowicz-Szczytt, 1991; Seiji Yamasaki et al., 2005). Similar to maize, it is the arrest of stamen or pistil development in initially bisexual flowers that leads to the development of unisexual flowers in cucumber (Atsmon & Galun, 1962; Malepszy & Niemirowicz-Szczytt, 1991). Furthermore, sex in cucumber is determined through the interplay of phytohormones, environmental factors, and genetic factors. In monoecious varieties of cucumber, sex determination tends to change as one moves along the stems. Lower nodes tend to produce male flowers, middle nodes produce both male and female flowers, and upper nodes tend to produce female flowers (Galun, 1961; Perl-Treves & Rajagopalan, 2006). Floral buds are bisexual until selective developmental arrest of either stamens or pistils results in unisexual flowers (or in the case of hermaphroditic flowers, the staminate and pistillate primordia continue to develop). In both male and female flowers, the spore-bearing parts of sexual organs are those that developmentally arrested. Specifically, the ovary never develops in male flowers, and the development of

the primordial anther is arrested in female flowers (S. L. Bai et al., 2004; Galun, 1961; Hao et al., 2003). The developmental arrest of these organs is based on location of the organs within the flower, rather then sexual identity of the organs (Kater, Franken, Carney, Colombo, & Angenent, 2001).

Several major genes affecting sex determination have been described, affecting both unisexual flower sex and spatial distribution. These genes are: the semi-dominant *F/f* gene, which controls femaleness, and affects the sex gradient observed on the plants; the *A/a* gene, which is epistatic to F and increases maleness; and the *M/m* gene, which determines whether flowers are unisexual or bisexual, and acts locally on individual buds that will develop an ovary (Galun, 1961; Kubicki, 1969a, 1969b, 1969c; Perl-Treves, 1999; R. W. Robinson, Munger, Whitaker, & Bohn, 1976). The M gene suppresses stamen development while the F gene shifts the femaleness downward in the plant by causing a higher levels of ethylene. Differing combinations of the *M, F*, and *A* loci lead to the wide variety of sexual phenotypes that are observed (reviewed in (Perl-Treves & Rajagopalan, 2006; Seiji Yamasaki et al., 2005)).

In addition to the genetic factors previously mentioned, phytohormones are also implicated in sex determination in cucumber. GA and ethylene have been found to affect the sexual phenotype of cucumbers, with GA primarily promoting maleness and ethylene, auxin, ABA, and cytokinin promoting femaleness (reviewed in (Perl-Treves, 1999; Seiji Yamasaki et al., 2005)). Additionally, the *M* and the *F* genes were found to encode ACC (1-aminocyclopropane-1-carboxylate) synthase genes, which are known to be the rate-limiting enzymes in the ethylene biosynthesis pathway (S. N. Bai & Xu, 2013; Boualem et al., 2009; Knopf & Trebitsh, 2006; Z. Li et al., 2009; Mibus & Tatlioglu, 2004; S.

Yamasaki, Fujii, Matsuura, Mizusawa, & Takahashi, 2001), and it has also been proposed that auxin influences sex expression in cucumber through the induction of ethylene biosynthesis (reviewed in (Seiji Yamasaki et al., 2005)). A recent publication suggests that a cucumber *GAMYB* gene (*CsGAMYB1*) can also regulate sex expression in an ethylene-independent fashion, acting to induce male flower development and/or inhibit female flower development(Y. Zhang et al., 2014).

Overall, it is clear that a combination of genetic and environmental factors come into play in sex determination in cucumber. The variety of sexual phenotypes as well as the myriad of physiological studies performed on cucumber make cucumber an excellent plant in which to study sex determination. However much still needs to be understood, such as the precise mechanisms involved in sex determination of unisexual, as well as the ways in which phytohormones regulate sex determination.

1.2.2 The dioecious angiosperms

It is thought that dioecy is an apomorphy that has evolved more then 100 different times (Charlesworth, 2002). As the sex determining mechanisms in dioecious species are very diverse, it is impossible in this brief introduction to cover all the dioecious plants in which sex determination has been studied. For the purposes of this chapter, the discussion will focus on sex determination in the dioecious plant *Silene latifolia*, (known formerly as *Melandrium album*), which is the dioecious angiosperm in which sex determination has been studied most extensively thus far. *Silene* is in the Caryophyllaceae family and phylogenetics has suggested that dioecy has arisen two separate times in this genus (Charlesworth, 2002; Desfeux, Maurice, Henry, Lejeune, &

Gouyon, 1996). This, along with the recent evolution of the Silene sex chromosomes, make Silene a particularly useful system for studying the evolution of sex chromosomes in that one can study the evolution of sex chromosomes in a time-course manner using various species in the *Silene* genus (reviewed in (Bernasconi et al., 2009)). Sex determination is diverse in *Silene* species; a number of species are dioecious with sex chromosomes; a number of species are not dioecious and do not have sex chromosomes; and one species (*Silene otitis*) is dioecious but lacks sex chromosomes (Filatov, 2005b). Male and female flowers form through the developmental arrest of anthers and gynoecium in female and male flowers respectively. Specifically, in female flowers the anthers are arrested in an early stage of sporogenesis and, as a result, the stamens are stunted (Farbos, Oliveira, Negrutiu, & Mouras, 1997). In male flowers, the stamens and anthers develop normally, while carpel initiation is prevented, and a functional pistil never develops (Farbos et al., 1997; Farbos et al., 1999; Grant, Hunkirchen, & Heinz, 1994).

Sexual phenotype in *Silene latifolia* is determined by morphologically distinct sex chromosomes (Westergaard, 1940, 1946). Silene has an XY system with XX female and XY male plants (Westergaard, 1940, 1946). The Y chromosome must lack certain essential genes, as YY plants are inviable (Ye et al., 1990). Genomic sequence and genetic mapping (Filatov, 2005a), as well as the fact that both hermaphroditic and dioecious species of Silene have the same number of chromosomes (2N=24), suggest that the sex chromosomes in dioecious species of Silene evolved from autosomes (Lebel-Hardenack, Hauser, Law, Schmid, & Grant, 2002; Moneger, Barbacar, & Negrutiu, 2000), likely in the last 10MYA (Filatov, 2005a). A number of cytological and mutagenesis

studies have been performed to elucidate the structure and function of the X and Y chromosomes.

In order to identify potential Y-linked mutations affecting stamen-promoting functions, irradiation of pollen and subsequent phenotypic screening and selection of asexual F_1 plants led to the identification of asexual (*asx*) mutants. These mutants were the result of deletion mutations on the Y chromosome, and display disrupted early stamen differentiation, at a developmentally identical stage to that at which stamen differentiation is arrested in wild-type female flowers. The alteration of phenotype seen in XY plants means that the deleted area responsible for early stamen differentiation does not have a functional counterpart at another location in the genome (Farbos et al., 1999). Hermaphroditic mutants, termed *bisexua* (*bsx*), resulted from two different types of mutations: those on an autosome, and those on the Y chromosome, with the strongest carpel suppressing locus residing on the Y chromosome (Lardon, Georgiev, Aghmir, Le Merrer, & Negrutiu, 1999). The *asx* mutants likely have a mutation in a gene(s) that promotes male development, while the *bsx* mutants likely have a mutation in a gene(s) that suppresses female development.

Multiple sex-linked genes have been identified and cloned, many of which have sex-specific expression (Filatov, 2005b, Kaiser et al., 2009), though the function of these genes remains largely unknown. A number of genes proposed to be involved in sex determination have also been discovered on autosomes, including orthologs of several ABC genes involved in floral development and organ identity in Arabidopsis (Koizumi et al., 2010; Zluvova, Nicolas, Berger, Negrutiu, & Moneger, 2006).

Further work in Silene can investigate the mechanisms responsible for sex determination and the development of dioecy in Silene. XX sex determining mutants have yet to be generated. Additionally, more work needs to be done to identify genes that are involved in sex determination, as little is currently known about the genes that determine sex or the molecular processes involved. With the advent of Next Generation Sequencing, identification of sex-linked and sex determination genes will no doubt proceed much faster. Already transcriptome sequencing has led to the discovery of many previously unidentified fully sex-linked and partially sex-linked genes (Bergero & Charlesworth, 2011; Bergero, Qiu, Forrest, Borthwick, & Charlesworth, 2013). These sex-linked genes, particularly those with homologs on both X and Y chromosomes, provide a valuable resource for studying the evolution of sex chromosomes (Bergero et al., 2013).

1.3 Sex determination in Bryophytes

The bryophytes are the lineage of plants that encompass the liverworts, hornworts, and mosses. In bryophytes, unlike in vascular plants, the haploid gametophyte is the dominant generation of the life cycle; the diploid sporophyte is dependent on and much smaller than the gametophyte. Liverworts, hornworts, and mosses all have some species which are homothallic (in which the gametophytes produce both egg and sperm producing gametangia), and have other species that are heterothallic (in which gametophytes produce either egg or sperm producing gametangia which are not on the same gametophytes and are thus unisexual) (G. M. Smith, 1955). All bryophytes are homosporous, producing only one type of spore. The first discovery of sex chromosomes

in plants was in the liverwort *Sphaerocarpus donnellii* (C.E. Allen, 1917; Charles E. Allen, 1919). Since then it has been shown that, in many species of heterothallic Bryophytes, sex is determined through sex chromosomes, making these Bryophytes the only known homosporous plants in which sex is determined through sex chromosomes (G. M. Smith, 1955).

Historically, bryophyte sex determining mechanisms have been most extensively studied in the heterothallic liverwort species *Marchantia polymorpha*, though some recent studies have focused on the model bryophyte *Physcomitrella patens*. Male and female Marchantia gametophtyes look nearly identical, with the exception of their reproductive structures. Female gametophytes bear archegoniophores, which produce egg-forming archegonia, and male gametophytes bear antheridiophores, which produce sperm-forming antheridia. The sex of Marchantia gametophytes is determined by heteromorphic sex chromosomes, with male gametophytes possessing small Y chromosomes and female gametophytes possessing larger X chromosomes (Lorbeer, 1934).

In contrast to Marchantia, *Physcomitrella patens* is a monoecious moss, with both male and female gametangia forming on the same gametophyte (Schaefer & Zryd, 2001). Studies on Physcomitrella have shown parallels in sex determination between bryophytes and vascular plants. A study was conducted to characterize the biological role of GAMYBs in Physcomitrella, an organism that lacks the GA perception and signal transduction pathways seen in higher vascular plants (Hirano et al., 2007). In angiosperms, GAs are known to modulate aspects of reproductive development such as floral organ formation and pollen development through the action of GAMYB

transcription factors (Aya et al., 2009; Gocal et al., 1999; Gocal et al., 2001; Kaneko et al., 2004). The results show GAMYBs to be necessary for both the initiation of male organ formation and for the suppression of female organ formation in Physcomitrella. Ultimately, the function of GAMYBs was found to be conserved between bryophytes and higher plants (Aya et al., 2011).

1.4 Sex determination in homosporous ferns

The following section on sex determination in homosporous ferns is a published review in *Frontiers in Plant Biology*, titled "Reproduction and the pheromonal regulation of sex type in fern gametophytes", and was authored by Nadia M. Atallah and Jo Ann Banks.

1.4.1 Introduction

The fern life cycle, illustrated in Figure 1.3, features two distinct body types: the large diploid sporophyte and the tiny haploid gametophyte. From a reproduction point of view, the sole function of the sporophyte is to produce then release haploid spores, while the gametophyte, which grows from a spore, functions to produce the gametes. Some ferns, like all angiosperms, are heterosporous and produce both mega- and microspores that are destined to develop as female and male gametophytes, respectively. Most ferns species are homosporous and produce only one type of spore. While textbook drawings of homosporous fern gametophytes typically show a heart-shaped hermaphrodite, fern gametophytes can be male, female, male then female, female then male, hermaphroditic or asexual, depending on the species. In this review we highlight old and recent studies

that have revealed the fascinating cross-talk that occurs between neighboring gametophytes in determining what their sexual phenotype will be.

1.4.2 Asexual reproduction in fern gametophytes

In addition to reproducing sexually, there are many examples of fern gametophytes that circumvent sex and reproduce asexually. The most common type of asexual reproduction is apogamy, whereby a sporophyte plant develops from a gametophyte without fertilization, similar to apomixis in angiosperms. In naturally occurring apogamous species, the viable spores produced by the sporophyte have the same chromosome number as the sporophyte (Walker, 1962, 1979). Obligate apogamy is associated with species of ferns that produce no or only one type of gametangia; because water is required for the flagellated sperm to swim to the egg in ferns, apogamous species are typically found in dry habitats where water is limiting (White, 1979). Apogamy also can be artificially induced in many ferns by adding sucrose to the culture media in which gametophytes are grown (White, 1979; Whittier & Steeves, 1962). By optimizing the conditions for inducing apogamy in *Ceratopteris richardii* gametophytes, a recent study has established *C. richardii* as a useful experimental system for studying this phenomenon (A.R. Cordle, Irish, & Cheng, 2007). Induced apogamous sporophytes of *C. richardii* have features typical of the sporophyte, including stomata, vascular tissue and scale-like ramenta; however, they are abnormal compared to sexually-derived diploid sporophytes, which could be a consequence of being haploid. To better understand how sucrose promotes the development of a sporophyte from cells of the gametophyte, the same researchers identified 170 genes whose expression is up-regulated during the period

of apogamy commitment. Many of them are associated with stress and metabolism or are homologs of genes preferentially expressed in seed and flower tissues (A. R. Cordle, Irish, & Cheng, 2012). Understanding apogamy, coupled with studies of apospory in *C. richardii*, where diploid gametophytes develop from cells of sporophyte leaves without meiosis (DeYoung, Weber, Hass, & Banks, 1997), should provide useful insights into genes and molecular mechanisms that regulate the alternation of gametophyte and sporophyte generations in ferns in the absence of meiosis and fertilization.

A second form of asexual reproduction in homosporous ferns involves vegetative propagation of the gametophyte. While relatively rare, such gametophytes typically do not produce sex organs. The fern *Vittaria appalachiana*, for example, is only known from its gametophytes (Farrar & Mickel, 1991). Each gametophyte forms vegetative buds, or gemmae, that allow gametophytes to multiply and form mats in dark, moist cavities and rock shelters in the Appalachian Mountains. While the origin of *V. appalachiana* (is it a recent hybrid or ancient relict?) and why it is unable to form sporophytes are unknown at this time, its persistent gametophyte suggest that fern gametophytes, like bryophyte gametophytes, can persist and thrive for very long periods of time.

1.4.3 Sexual reproduction

Most homosporous ferns that reproduce sexually ultimately form hermaphroditic gametophytes that have antheridia and archegonia. While hermaphroditism increases the probability that a single gametophyte will reproduce, self-fertilization of a hermaphrodite (which is genetically similar to a doubled haploid in angiosperms) results in a completely

homozygous sporophyte. Given that this absolute inbreeding could have negative consequences to the individual and reduce genetic variation in populations, it is not surprising that homosporous ferns have evolved mechanisms to promote outcrossing. One such mechanism that is common to many species of ferns involves the pheromonal regulation of sexual identity, where the sexual phenotype of an individual gametophyte depends on its social environment.

1.4.4 One genotype-two or more phenotypes

In the late 1800's, botanists began noting that fern gametophytes are often sexually dimorphic, with larger gametophytes bearing archegonia and smaller gametophytes bearing antheridia (Prantl, 1881; Yin & Quinn, 1995). The size difference between them was attributed to the presence or absence of a meristem, with females or hermaphrodites being "meristic" (with a meristem) and males "ameristic" (without a meristem). In a major discovery, Döpp noted that the medium harvested from cultures of *Pteridium aquilinum* gametophytes contained a pheromone that promoted the development of males in juvenile gametophytes (Döpp, 1950a); this pheromone is referred to as antheridiogen. Antheridiogens or antheridiogen responses have since been identified in over 20 species of ferns (Jimenez, Quintanilla, Pajaron, & Pangua, 2008; Kurumatani et al., 2001; Yamane, 1998a).

Much of what is known about the biology of antheridiogen responses can be attributed to studies by Näf and Schraudolf during the 1950s and 1960s (reviewed in (Naf, 1979; Näf, 1959). This response is illustrated here for the fern *Ceratopteris richardii*, originally characterized by Hickok (Hickok, Warne, & Fribourg, 1995). In this species,

an individual spore always develops as a relatively large hermaphrodite (Fig. 1.4A) that produces egg-forming archegonia (Fig. 1.4B), sperm-forming antheridia, and a multicellular lateral meristem. The hermaphrodite also secretes antheridiogen, or A_{CE} (for antheridiogen Ceratopteris) into its surroundings. If the hermaphrodite is removed then replaced with a genetically identical spore, the new spore will develop as an ameristic male gametophyte (Fig. 1.4C) with many antheridia (Fig. 1.4D) in response to A_{CE} secreted by the hermaphrodite. In a population of spores, spores that germinate first become hermaphrodites that secrete A_{CE}, while slower-growing members of the population become male in response to the secreted A_{CE}. In comparison to chromosomal based sex determination, this mechanism of sex-determination is unusual because it allows the ratio of males to hermaphrodites to vary depending on population size and density and it is inherently flexible rather than fixed.

Typical of other ferns, a *C. richardii* gametophyte is able to respond to A_{CE} for a limited period of time, prior to the establishment of a lateral meristem. The lateral meristem not only confers indeterminate growth to the gametophyte, but its formation coincides with a loss in ability to respond to A_{CE} as well as the secretion of A_{CE}. Archegonia invariably initiate close to the meristem notch of the hermaphrodite, well after the lateral meristem is well developed. While the hermaphroditic program of expression cannot be reversed, the male program of expression is reversible. Cells of the male gametophyte prothallus, when transferred to media lacking A_{CE}, will divide to ultimately form one or more new hermaphroditic prothalli (Fig. 1.4E). Antheridiogen thus serves multiple functions in male gametophyte development: it represses divisions of the prothallus that establish the lateral meristem; it promotes the rapid differentiation of

antheridia; it represses its own biosynthesis; and it serves to maintain in the gametophyte an ability to respond to itself.

All of the antheridiogens that have been structurally characterized from ferns are gibberellins (GAs) (Furber, Mander, Nester, Takahashi, & Yamane, 1989; Takeno et al., 1989; Yamane, 1998b; Yamane, Nohara, Takahashi, & Schraudolf, 1987a). Although the structure of A_{CE} is unknown, GA biosynthetic inhibitors reduce the proportion of males in a population of *C. richardii* gametophytes suggesting that A_{CE} and GA share a common biosynthetic pathway (T. R. Warne & Hickok, 1989). ABA, a known antagonist of GA responses in angiosperms, completely blocks the A_{CE} response in *C. richardii*, also indicating that A_{CE} is likely a GA (Hickok, 1983).

1.4.5 The sex-determining pathway in Ceratopteris

Most recent studies aimed at understanding how antheridiogen determines the sex of the gametophyte have focused on two species of homosporous ferns: *C. richardii* and *Lygodium japonicum*. *Ceratopteris richardii* is a semi-tropical, annual species and is useful as a genetic system for many reasons. Large numbers of single-celled, haploid spores (typically 10^6) can be mutagenized and mutants identified within two weeks after mutagenesis. Gametophytes can be dissected and regrown, making it possible to simultaneously self-fertilize and out-cross a single mutant gametophyte. Because self-fertilization of a gametophyte results in a completely homozygous sporophyte that produces $>10^7$ spores within a six-month period, suppressor mutants are also easy to generate. Because *C. richardii* gametophytes are sexually dimorphic, mutations affecting the sex of the gametophyte are especially easy to identify (Banks, 1994b, 1997a, 1997d;

Chun & Hickok, 1992; Eberle & Banks, 1996; Hickok, 1977, 1985; Hickok & Schwarz, 1989; Hickok, Scott, & Warne, 1985; Hickok, Vogelien, & Warne, 1991; Renzaglia, Wood, Rupp, & Hickok, 2004; Scott & Hickok, 1991; Strain, Hass, & Banks, 2001; Vaughn, Hickok, Warne, & Farrow, 1990; T. R. Warne & Hickok, 1986; T. R. Warne, Hickok, & Scott, 1988). Over 70 mutants affecting sex determination have been characterized, most falling into three major phenotypic groups: the *hermaphroditic* (*her*) mutants, which are hermaphroditic in the presence or absence of A_{CE}, the *transformer* (*tra*) mutants, which are male in the presence or absence of A_{CE}, and the *femininization* (*fem*) mutants, which are female in the presence or absence of A_{CE} and produce no antheridia. Through test of epistasis (i.e., comparing mutant phenotypes of single and various combinations of double and triple mutants), a genetic model of the sex determination pathway has been developed and is illustrated in Fig. 1.5 (Banks, 1997a, 1997d; Eberle & Banks, 1996; Strain et al., 2001). This pathway reveals that there are two major regulators of sex: *TRA*, which is necessary for lateral meristem and archegonia development (female traits), and *FEM*, which is necessary for antheridia development (the male trait). *FEM* and *TRA* negatively regulate each other such that only one can be expressed in the gametophyte. What determines whether *FEM* or *TRA* is expressed in the gametophyte is A_{CE}. A_{CE} activates the *HERs*, which, in turn, repress *TRA*. Because *TRA* cannot repress *FEM*, *FEM* is expressed and the gametophyte develops as a male. In the absence of A_{CE}, *HER* is not active and is thus unable to repress *TRA*. *TRA* promotes the development of a gametophyte with female traits and represses the development of antheridia by repressing the *FEM* gene that promotes male development. Additional genetic experiments have revealed that the repression of *FEM* by *TRA* and of *TRA* by

FEM is indirect and involves other genes (Strain et al., 2001). What is remarkable about this pathway is that it is inherently flexible, which is consistent with what is understood about sex determination in this species by A_{CE}. This "battle of the sexes"—deciding whether to be male or female—depends on which of the two major regulatory sex genes prevails in the young gametophyte, a decision that is ultimately determined by the presence or absence A_{CE}.

While this model explains how male and female gametophyte identities are determined, it does not explain the hermaphrodite. One possibility is that in certain cells of the hermaphrodite, the activities of *FEM* and *TRA* are reversed, allowing *FEM* to be expressed in cells that will eventually differentiate as antheridia. Testing this and other possibilities will require the cloning of the sex-determining genes and assessing their temporal and spatial patterns of expression in the developing hermaphrodite.

The sex-determining pathway in *C. richardii* is remarkable in its resemblance to the GA signaling pathway in angiosperms (Sun, 2011), as illustrated in Fig. 1.5. In Arabidopsis, GA is bound by its receptor GIBBERELLIN INSENSITIVE DWARF1 (GID1). The GA-GID1 complex triggers the rapid proteolysis of one or more DELLA proteins, a type of GRAS family transcription factors that are ultimately responsible for repressing GA responses (Sun, 2011). Proteolysis of DELLA requires GID1 and the specific F-box protein SLEEPY1 (SLY1), which promotes poly-ubiquitination of DELLA by the $SCR^{SLY1/GID2}$ complex and results in its degradation by the 26S proteasome. Since DELLA acts as a repressor of GA responses, its GA-induced degradation results in a GA response. While targets of DELLA repression have been identified (Fleet & Sun, 2005), in the case of barley seed germination (which requires GA), DELLA directly or indirectly

represses *GAMYB*, a transcription factor that promotes a-amylase expression in germinating barley seeds (Gubler, Kalla, Roberts, & Jacobsen, 1995; Gubler et al., 1999). Based on the similarities between the GA signaling pathway in angiosperms and the sex determination pathway in *C. richardii*, we hypothesize that the *HER* genes in *C. richardii* encode GID1 and SLY1, that *TRA* encodes a DELLA protein, and that *FEM* encodes a GAMYB-like protein. These hypotheses can be tested by sequencing these candidate genes from mutant and wild-type plants and by knocking-down their expression in the gametophyte by RNAi methods well established in *C. richardii* (Rutherford, Tanurdzic, Hasebe, & Banks, 2004b).

1.4.6 Antheridiogen biosynthesis is split between young and older gametophytes in *Lygodium japonicum*

Lygodium japonicum is another homosporous fern species with an antheridiogen response. This species has the distinct advantage of having its antheridiogens structurally well characterized. Two different GAs have been identified as antheridiogens in this species, including GA_9 methyl ester (Yamane, Takahashi, Takeno, & Furuya, 1979) and GA_{73} methyl ester (Yamane et al., 1988). GA_{73} methyl ester is the most active antheridiogen and is able to induce antheridia formation at the incredibly low concentration of 10^{-15} M. To test the hypothesis that antheridiogen is synthesized through the GA biosynthetic pathway, *L. japonicum genes* related to five different GA synthesis genes, including *ent-copalyl diphosphate/ent-kaurene synthase* (*CPS/KS*), *ent-kaurenoic acid oxidase* (*KAO*), *kaurene oxidase* (*KO*), *GA 20-oxidase* (*GA20ox*) and *GA3-oxidase* (*GA3ox*), were identified and their expression patterns in developing

gametophytes investigated (Tanaka et al., 2014). Their expression patterns revealed that all but *GA3ox* were more highly expressed in older gametophytes that secrete antheridiogen, consistent with the expectation that antheridiogen biosynthesis genes are up-regulated in gametophytes that secrete it. *GA3ox* showed the opposite pattern of expression; i.e., it was more highly expressed in young gametophytes that did not secrete antheridiogen but were capable of responding to antheridiogen. To explore this further, the same authors assayed the effects of prohexadione, a GA3ox inhibitor, on antheridia formation in the presence of GA_4 (which has an OH group at the C3 position) or GA_9 methyl ester (which lacks the OH group at C3); both GA_9 and GA_4 induce antheridia formation by themselves. Whereas prohexadione plus GA_9 methyl ester inhibited antheridia formation, prohexadione plus GA_4 did not, demonstrating that C3 hydroxylation of antheridiogen is essential for inducing antheridia formation. In another series of experiments, the authors found that GA_9 methyl ester was converted to GA_9 in young gametophytes. Based on these and other results, a model was proposed whereby antheridiogen (GA_9 methyl ester) is synthesized via a GA biosynthetic pathway and secreted by older gametophytes. When it is taken up by younger gametophytes, the methyl ester is removed by a possible methyl esterase then hydroxylated at the C3 position by GA3ox to GA_4, where it is perceived and transduced by the GA signaling pathway in young gametophyte. Because GA_9 methyl ester is more hydrophobic and more efficiently taken up by gametophytes than GA_9, splitting the GA biosynthetic pathway between young and older gametophytes was proposed to enhance the sensitivity of young gametophytes to the secreted antheridiogen by their neighbors and, at the same

time, promote the activation of male traits once inside the young gametophyte (Tanaka et al., 2014).

In addition to characterizing antheridiogen biosynthesis in *L. japonicum*, Tanaka et al. also made two other important discoveries. They found that a *L. japonicum* DELLA protein was degraded in GA_4 and GA_9 methyl ester treated gametophytes, and that the *L. japonicum* GID1 and DELLA proteins could interact in a yeast –two hybrid assay, but only in the presence of GA_4 (and not GA_4 methyl ester or GA_9 methyl ester). All told, the results of these experiments were used to define a model of the antheridiogen response in *L. japonicum* that is remarkably similar to the pathways illustrated in Figure 1.5.

1.4.7 Future Directions

The elucidation of the antheridiogen biosynthetic and signaling pathways in ferns has only just begun and many questions regarding sex determination and sexual reproduction remain, many of which can be resolved by cloning all of the sex determining genes. Some of these questions are: To what extent are other hormones involved in sex determination? Is the split GA biosynthetic pathway in *L. japonicum* typical of other ferns? What is the relationship between the antheridiogen response in the gametophyte to GA responses in the sporophyte? Knowing that some mutations in *C. richardii* (e.g., *her* mutations) have no effect on the sporophyte while other mutations (e.g., *tra* mutations) severely affect the sporophyte suggest that at least some, but not all, genes are necessary in both generations. Is antheridiogen also involved in the developmental decision to produce mega- and micro-sporangia in heterosporous ferns?

From an evolutionary perspective, was the antheridiogen signaling and responses in the gametophyte co-opted during or important for the evolution of heterospory from homospory in ferns? Addressing these and other questions will lead to a more comprehensive understanding of sex determination in ferns, including an understanding of the molecular mechanisms at play.

1.5 Conclusion

Sex determination is a fundamental process in the development of many plants. Although the majority of plants are hermaphroditic, there are a considerable number of species that have separate sexes, including many economically important plants. Because the separation of sexes seems to have evolved hundreds of times, and thus the sex determination mechanisms employed in plants are broad, sex determination will need to be studied in a multitude of plant species to gain a comprehensive understanding of sex determination in plants. Gaining insight into sex determination mechanisms in a range of plant species and clades will also improve understanding of how heterospory evolved from homospory.

1.6 Purpose of Proposed Research

Ceratopteris richardii is an excellent system for studying sex determination in plants for a number of reasons. First, we know what determines sex, and also when sex is determined in Ceratopteris. The rapid life cycle of Ceratopteris and the fact that it is an exceptional genetic system add to the value of this system for understanding the intricacies of sex determination in plants, particularly in homosporous plants. As stated

previously, a number of sex determining mutants have been identified in Ceratopteris and a genetic sex determination pathway has been described using tests of epistasis (Banks, 1993, 1994b, 1997c; Strain et al., 2001). Unfortunately identification of these genes is not possible using traditional techniques due to the large genome size and lack of a reference genome in Ceratopteris and thus a Next-Generation sequencing approach was taken to obtain sequence information from Ceratopteris gametophytes and to identify potential sex-determining genes in Ceratopteris.

To assemble a reference transcriptome, identify genes potentially involved in sex determination in Ceratopteris, and assess the changes in the transcriptome over time during early gametophyte development, RNA-Seq and differential expression analyses were performed. It was hypothesized that using RNA-Seq, a Ceratopteris transcriptome could be assembled and differentially expressed genes could be identified between $+A_{CE}$ and $-A_{CE}$ conditions. Chapter 2 describes an RNA-Seq experiment that led to the assembly of the transcriptome of gametophytes at 4.5 DAI (days after inoculation). In this experiment gametophytes were treated or not treated with A_{CE} at 3 DAI, grown for an additional 1.5 days, RNA isolated and sequenced, and differentially expressed genes were identified between conditions. Chapter 3 details a time-course RNA-Seq experiment in which the transcriptomes of gametophytes at 0, 3, 3.5, 4.5, and 5.5 DAI were sequenced, assembled, and expression patterns across development identified. Concluding comments are given in Chapter 4, summarizing experimental results and providing information on experiments that are underway to test the hypotheses identified using the RNA-Seq experiments.

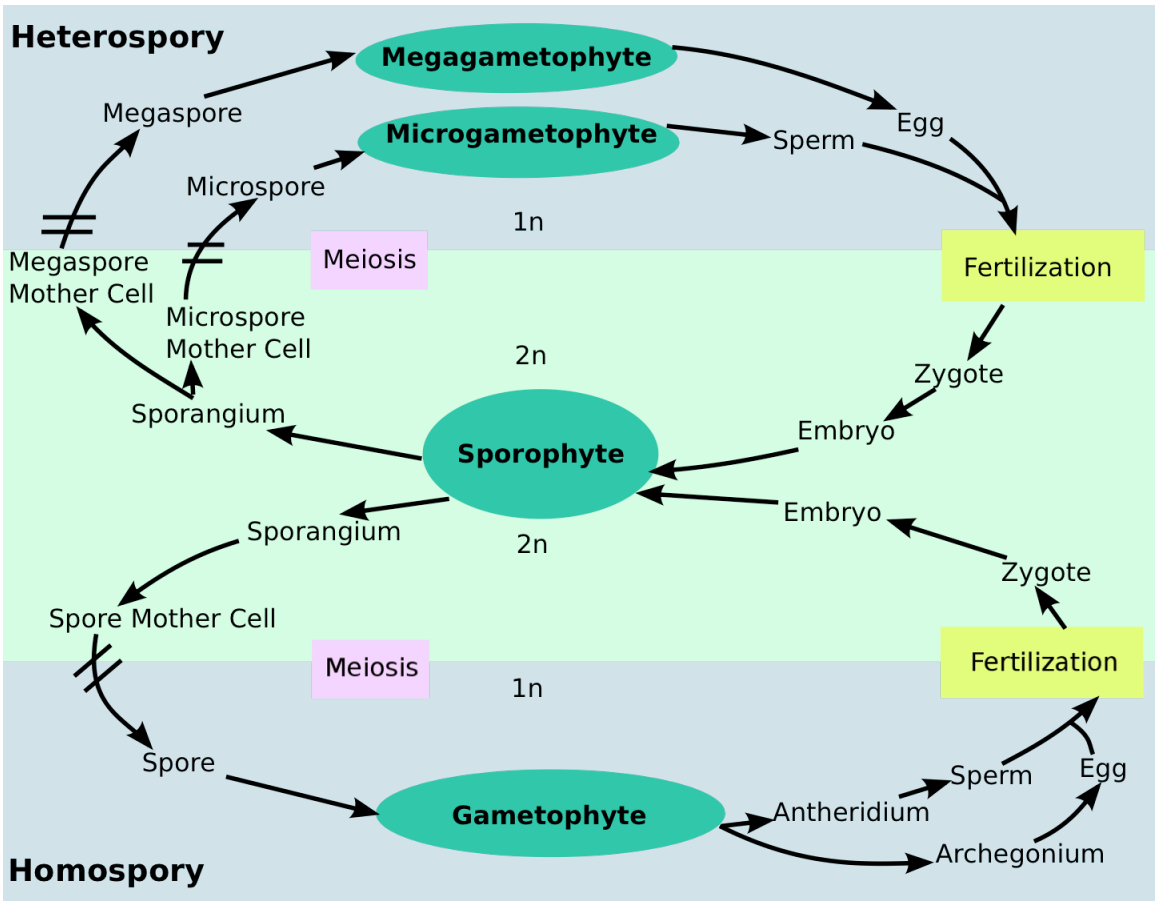


Figure 1.1. Homospory versus heterospory in plant life cycles. In heterospory, the sporophyte produces a sporangium that contains either megaspore mother cells or microspore mother cells, which undergo meiosis to produce megaspores and microspores, respectively. Megaspores then form the megagametophyte, which then produces egg cells whereas the microspores produce microgametophytes, which produce sperm. In homospory, the diploid sporophyte produces a sporangium, which contains the spore mother cells. Meiosis occurs, leading to production of haploid, sexually undetermined spores. These spores then germinate and grow into haploid gametophytes, the sexual stage of the life cycle. Gametophytes then produce sperm containing antheridia and egg-containing archegonia. In both heterospory and homospory, upon fertilization of the egg by sperm, a zygote is formed, which then develops into a diploid sporophyte. Blue sections of the figure indicate haploid stages of the cycle whereas the green section of the figure indicates diploid stages of the cycle.

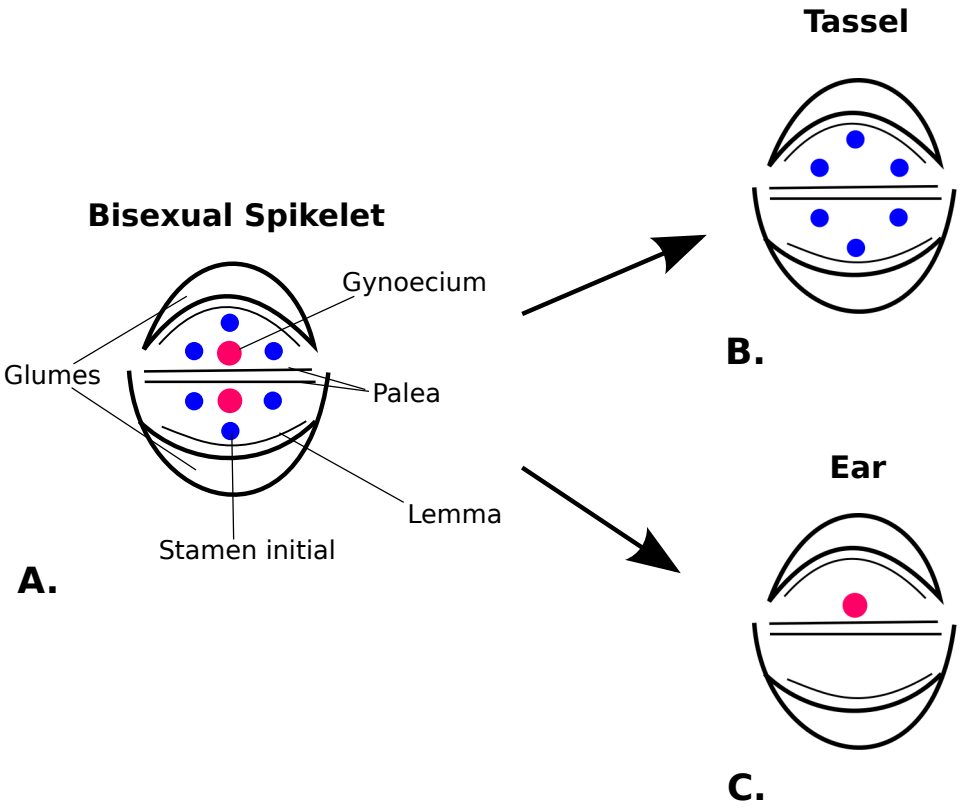


Figure 1.2. Floral diagrams of spikelet structure in maize. A. The bisexual stage of the maize spikelet, in which the tassel and ear florets are indistinguishable. Each spikelet consists of 2 florets, each with a lemma, palea, gynoecium, three stamen initials, and each subtended by a glume. B. In the tassel, which is destined to be male, the gynoecium in both florets are aborted. C. In the ear, which is destined to be female, the stamen primordia are aborted in both florets, as well as the gynoecium in the secondary floret.

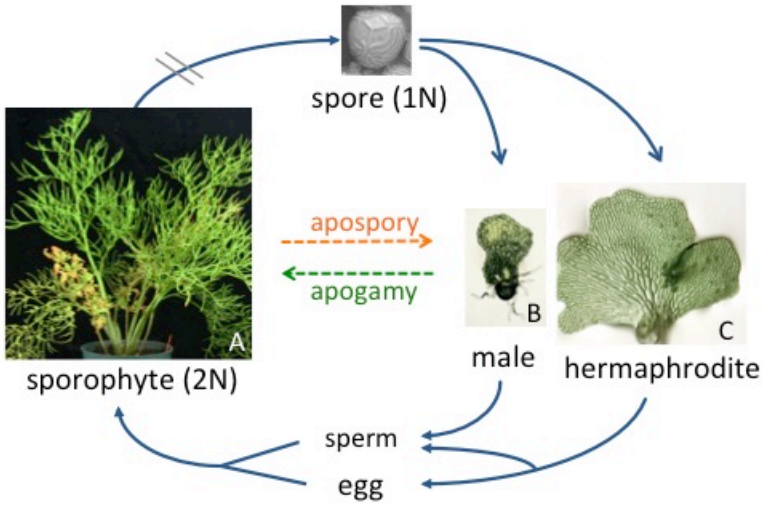


Figure 1.3. The *C. richardii* life cycle. Typical of all homosporous ferns, the diploid sporophyte produces sporangia on the abaxial surface of the fronds. Each sporangium contains haploid spores that are released from the sporophyte and, in the case of *C. richardii*, can remain dormant but viable for more than 50 years. Each spore germinates and develops as a male or hermaphroditic gametophyte depending on the presence or absence of antheridiogen. When mature, sperm are released and swim to the egg. The young sporophyte remains dependent on the gametophyte for a short period of time.

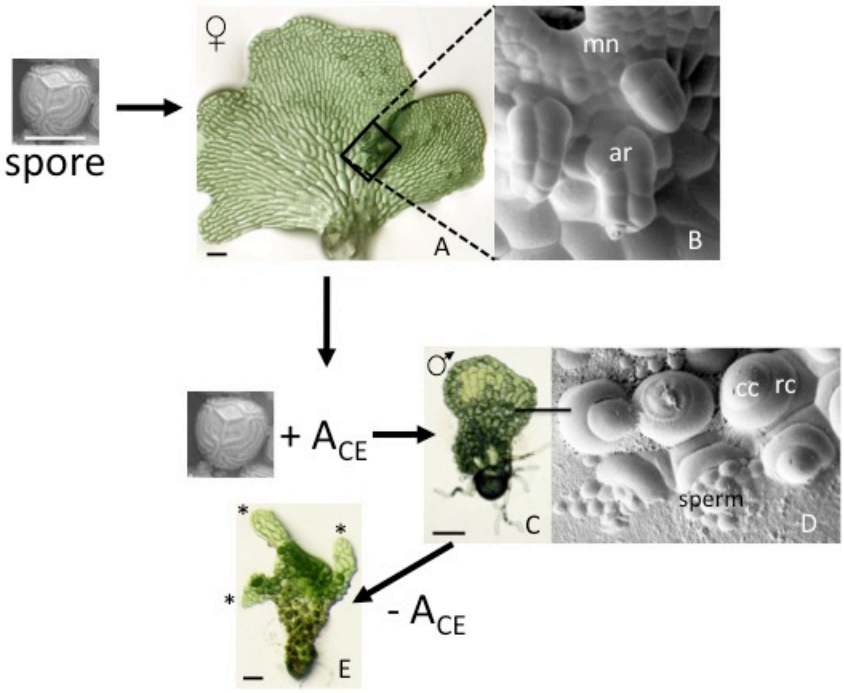


Figure 1.4. The antheridiogen response *in C. richardii*. A single spore always develops as a hermaphrodite when grown in the absence of A_{CE}. The hermaphrodite consists of a single sheet of cells with a distinct multicellular meristem that forms a meristem notch and multiple archegonia that develop adjacent to the meristem notch, which are highlighted in the SEM (boxed area of the hermaphrodite). Hermaphrodites secrete A_{CE}; in the presence of A_{CE}, spores develop as males. The male lacks a meristem and almost all cells differentiate as antheridia. The SEM shows six antheridia, each having a ring cell and a cap cell that pops open to release sperm. When a male gametophyte is transferred to media lacking A_{CE}, some cells divide and begin to form a hermaphroditic prothallus. The "switched" male shown is forming three such prothalli. mn: meristem notch; ar: archegonia; cc: cap cell; rc: ring cell.

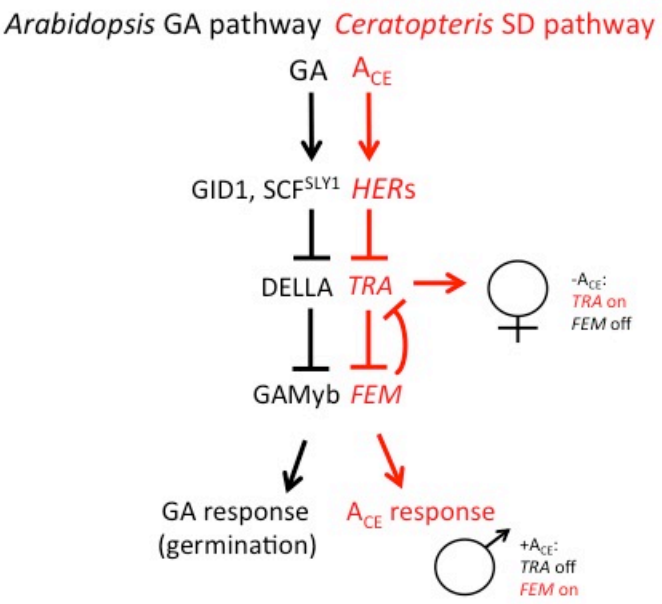


Figure 1.5. A comparison of the GA signaling pathway in angiosperms and the sex-determining (SD) pathway in *C. richardii*. The SD pathway in *C. richardii* is based solely on the epistatic interactions among sex-determining mutants but it is consistent with recent molecular and biochemical studies in the fern *L. japonicum*. T bars represent repressive events whereas arrows indicate activating events.

CHAPTER 2. SEX DETERMINATION AND TRANSCRIPTIONAL REPROGRAMMING OF *CERATOPTERIS RICHARDII* GAMETOPHYTES BY ANTHERIDIOGEN

2.1 Introduction

Ceratopteris richardii is a homosporous fern that produces a single type of haploid spore, with each spore having the potential to develop as either a free-living male or hermaphroditic gametophyte. In this and many other fern species, the sex of the gametophyte is determined by a pheromone called antheridiogen (Banks, 1999; T.R. Warne & Hickok, 1991), first discovered by Döpp in the fern *Pteridium aquilinum* (Döpp, 1950b). In the absence of A_{CE} (for **a**ntheridiogen **Ce**ratopteris), a Ceratopteris spore develops as a hermaphrodite that begins to secrete biologically detectable amounts of A_{CE} after losing the competence to respond to its male-inducing effects. In the presence of A_{CE}, a spore develops as a male gametophyte. Thus, in a population, spores that germinate first in the absence of A_{CE} develop as hermaphrodites that secrete A_{CE}, while spores that germinate later and in the presence of A_{CE} develop as males (Banks, 1997b; J. A. Banks, L. G. Hickok, & M. A. Webb, 1993c; T.R. Warne & Hickok, 1991). Although small (<3mm), male and hermaphroditic gametophytes are dimorphic and easily distinguished by size and shape at maturity. Each hermaphrodite forms a multicellular, lateral meristem that contributes to its heart-shaped appearance, with multiple egg-forming archegonia developing after the lateral meristem forms (Fig. 2.1G). The

development of this lateral meristem coincides with the loss of competence to respond to A_{CE} in the hermaphrodite as well as the production of A_{CE}. Male gametophytes never develop a lateral meristem and are much smaller than hermaphrodites (Fig. 2.1D), with nearly all cells of the male gametophyte terminally differentiating as antheridia. Based on these observations, A_{CE} has two primary functions in early gametophyte development: it suppresses the indeterminate growth by suppressing the divisions of the gametophyte that give rise to the lateral meristem in the hermaphrodite and promotes the rapid differentiation of antheridia in the male.

All antheridiogens that have been structurally characterized from ferns are gibberellins (GAs) (Furber et al., 1989; Takeno et al., 1989; Yamane, 1998b; Yamane et al., 1987a). Although the structure of A_{CE} is unknown, the GA biosynthetic inhibitors ancymidol, AMO-1618, and uniconazole-P reduce the proportion of males in a population of Ceratopteris gametophytes suggesting that A_{CE} and GA have a common biosynthetic pathway (T. R. Warne & Hickok, 1989). That ABA completely blocks the A_{CE} response in Ceratopteris is also consistent with A_{CE} being a GA (Hickok, 1983).

To understand how A_{CE} determines the sex of the Ceratopteris gametophyte by suppressing female traits (meristem and archegonia) and promoting male traits (antheridia), mutations affecting the sex of the gametophyte have been characterized and used to develop a genetic model of the sex-determining pathway (Banks, 1994b, 1997d; Eberle & Banks, 1996; Strain et al., 2001). Cloning these genes is challenging because of the large genome size of *C. richardii* (~9Gb) (J. Banks unpub. obs.) and the lack of a reference genome sequence for any fern. An alternative approach to identifying potential sex-determining and differentiation genes involves *de novo* transcriptome assembly using

RNA-seq, which provides a means to perform sensitive gene expression studies in organisms that do not have a reference genome (Grabherr et al., 2011b; Robertson et al., 2010; Schulz, Zerbino, Vingron, & Birney, 2012). The Ceratopteris gametophyte is well-suited to this approach for identifying genes involved in sex determination and differentiation for several reasons. Gametophyte development is independent of the sporophyte and gametophytes are easy to grow and manipulate. The sex of the Ceratopteris gametophyte is determined during a brief period of time, about 3.5-4.5 days after spore inoculation, as the single-cell spore nucleus begins to divide (Banks et al., 1993c). At this time, the gametophyte consists of three or fewer cells and is not a complex tissue that could confound the interpretation of RNA-seq results. Finally, a hermaphrodite can be easily self-fertilized, leading to a homozygous sporophyte (similar to a doubled haploid) that produces millions of genetically identical spores, thereby avoiding potential problems associated with heterozygosity in RNA-seq experiments.

Here we describe the *de novo* assembly of the transcriptome of young Ceratopteris gametophytes, identify genes whose expression differs between gametophytes as their sex is being determined by A_{CE}, and identify candidate sex-determining genes known only by their mutant phenotypes. The functions of candidate genes can be tested in the future, either by knocking-down gene expression transiently by RNAi in the gametophyte (Rutherford, Tanurdzic, Hasebe, & Banks, 2004a), or altering gene expression in stably transformed sporophyte and gametophyte plants (Plackett, Huang, Sanders, & Langdale, 2014).

2.2 Materials and methods

2.2.1 Plants and growth conditions

The origin of Hn-n, the wild-type strain of *Ceratopteris richardii* used in this study, is described in (L.G. Hickok, T. R. Warne, & M. K. Slocum, 1987). The conditions for spore sterilization and gametophyte culture are as previously described (Banks, 1994b). Medium used to culture gametophytes in the absence of exogenous A_{CE} is as described in (Banks et al., 1993c) and is referred to as fern medium, or FM. A_{CE} was obtained as a crude aqueous filtrate from media previously supporting gametophyte growth in FM as described in (Banks et al., 1993c) and is referred to as conditioned FM (CFM). Scanning electron micrographs (SEMs) were performed on a FEI NOVA nanoSEM on samples prepared as previously described (Banks, 1994b).

For both RNA-seq and qRT-PCR, spores were grown aseptically in liquid FM at 28°C in a growth chamber, shaken at 105rpm, and at a density of 1g spores/L. Three days after spore inoculation, gametophytes were filtered from media; 1/6 of the spores were added to each of three flasks containing 200 mL sterile FM, which is the $-A_{CE}$ treatment, and 1/6 were added to each of three flasks containing 200 mL sterile CFM, which is the $+A_{CE}$ treatment. After 36 hours, gametophytes were vacuum filtered from media and frozen in N_2(l). Tissue was subsequently stored at -80°C.

2.2.2 Library preparation and sequencing

Frozen tissue was ground under N_2(l) for 30 minutes and total RNA extracted using the RNeasy Plant Mini Kit (Qiagen, CA). The TruSeq kit (Illumina, CA) was used to select poly-adenylated mRNA and prepare six non-directional libraries for sequencing.

Libraries were sequenced on an Illumina HiSeq2000 platform using paired-end technology.

2.2.3 Transcriptome assembly and quality control

DeconSeq version 0.4.1 was run on each of the FASTQ read files to remove reads aligning to bacterial, viral, rRNA, mitochondrial RNA, and chloroplast DNA (Schmieder & Edwards, 2011b; Schmieder, Lim, & Edwards, 2012b). An identity threshold of 75 and a coverage value of 50 were used. The program clean_adapter.pl version 1.4 (Gribskov, pers. comm.) was used to remove Illumina adapter sequences. The program Trimmomatic version 0.22 was used to trim reads based on quality score (Lohse et al., 2012a). Reads that were under 30 bases long post-trimming were removed. Local base trimming was performed to trim internal bases with poor quality scores. A sliding window of 4 bases was used across reads, trimming those whose average Phred quality score was less than 13. This allows one base to be of low quality without discarding the read, however it does not allow two bases to be of low quality within the window of 4. The default in Trimmomatic is to trim bases at the beginnings or ends of reads with Phred quality score less than 3. However to be slightly more conservative a cutoff of 7 was used. Reads were next assembled using the *de novo* transcriptome assembler Trinity (release 2012-06-08), with a minimum contig length cutoff of 150. Trinity utilized a fixed k-mer size of 25 to identify read overlaps (Grabherr et al., 2011b). Trinity output assigns predicted transcripts a three-part name as a result of the assembly algorithm. The program Assembly Stats in the iPlant Discovery environment was utilized to obtain basic assembly statistics (Earl et al., 2011; Goff et al., 2011). R code, custom scripts and commands used in the analyses of this data are included in Appendix A.

2.2.4 Differential expression analysis

The program cmpfastq-pe.pl (Newhouse & To, 2010) was run on FASTQ files to separate reads into paired and unpaired reads. Paired reads were aligned to the assembled transcriptome using RSEM (Grabherr et al., 2011b; B. Li & Dewey, 2011; B. Li, Ruotti, Stewart, Thomson, & Dewey, 2010). RSEM was run with components representing the gene level. Only the transcripts with at least one read aligned in at least one of six samples were used as an input. The programs edgeR v. 3.0.8 (M. D. Robinson, McCarthy, & Smyth, 2010), DESeq v. 1.10.1 (Anders & Huber, 2010), and EBSeq v. 1.1.4 (Leng et al., 2013) were used to identify differentially expressed genes at a Benjamini-Hochberg corrected FDR (Benjamini, Drai, Elmer, Kafkafi, & Golani, 2001) of q=0.01. In edgeR, dispersion was estimated as tagwise dispersion. An additional fold-change cutoff of 2 was applied in selecting differentially expressed genes.

2.2.5 Annotation and assembly validation

Protein-encoding, differentially expressed genes were annotated using the Trinotate workflow (Ashburner et al., 2000; Finn, Clements, & Eddy, 2011; Grabherr et al., 2011b; Kanehisa, Goto, Sato, Furumichi, & Tanabe, 2012) using the version released on 2013-02-25, and a 50 amino acid minimum cutoff for annotated ORFs. BLAST2GO (Aparicio et al., 2006; Conesa & Gotz, 2008; Conesa et al., 2005; Gotz et al., 2008) was run and multilevel pie charts made for all predicted transcripts with read support. For the BLAST2GO annotation of predicted transcripts, sequence number cutoffs of 2000 for biological process, 500 for cellular component, and 500 for molecular function GO terms were used. For the annotation of differentially expressed genes, sequence number cutoffs

of 55 for biological process, 10 for cellular component, and 13 for molecular function GO terms were used. In hand annotating each predicted transcript, a BLASTx search, using the Ceratopteris gene as query, followed by a reciprocal tBLASTn search against the Ceratopteris transcriptome, was performed for each differentially expressed gene. With the exception of transposon-derived transcripts and putative cytochrome P450 genes, a Ceratopteris gene was considered to be a similar to a known gene if it gave a reciprocal best BLASTx hit (E-values $<2x10^{-30}$) and if it was identified as orthologous using the program OrthologID (http://nypg.bio.nyu.edu/orthologid/), which automates gene orthology determination within a character-based phylogenetic framework (Chiu et al., 2006).

To assess the quality of the Ceratopteris Trinity assembly, the Ceratopteris Sanger-generated ESTs available in GenBank were used to blast the entire Ceratopteris transcriptome assembly using BLASTn.

2.2.6 Expression analysis validation

Total RNA was reverse transcribed into single-stranded cDNA using the Tetro cDNA Synthesis Kit (Bioline, MA). Approximately 3 ng cDNA was used as template for each qRT-PCR reaction, performed using the SYBR green PCR Master Mix from Applied Biosystems and the StepOne Real-Time PCR System (Applied Biosystems, NY). All oligonucleotide primers were used at a 900nM concentration. PCR conditions were: one cycle of 20 minutes at 95°C, 40 cycles of 3 seconds at 95°C and 30 seconds at 60°C. Melt curves (15 seconds at 95°C, 60 seconds at 60°C, and 15 seconds at 95°C) were performed and only those reactions producing a single Tm peak used. Three biological

replicates of both $+A_{CE}$ and $-A_{CE}$ samples were performed for each template and three technical replicates were performed for each sample. Measurements were normalized to the amount of CrEF1α (GenBank accession number BE642078) transcript in the samples. Reactions without template added served as the negative control. The ΔCt method was used in calculating relative fold changes (Livak & Schmittgen, 2001). The primer sequences used are listed in Table 2.1.

2.3 Results and discussion

2.3.1 Gametophyte morphology

To identify the genes that are differentially expressed as sex is determined by A_{CE}, 4.5d old gametophytes were grown in media without A_{CE} or with A_{CE} present between 3 and 4.5d after spore inoculation. If a gametophyte is not continuously exposed to A_{CE} between 3-4.5d it will develop as a hermaphrodite (Fig. 2.1G) and if exposed continuously to A_{CE} during the same period of time, it will develop as a male (Fig. 2.1D). The Ceratopteris spore swells until day 4 when the spore wall opens at its trilete markings, shown in Figure 2.1A. At 4.5d when gametophytes were harvested for RNA-seq, the protonema consisted of at most three cells with rhizoids (Figs. 2.1B and 2.1E). Morphological differences between gametophytes grown in the presence or absence of A_{CE} were not apparent until 6d (Figs. 2.1C and 2.1F), at which time antheridia and a lateral meristem begin to differentiate in males and hermaphrodites, respectively.

2.3.2 RNA-seq and *de novo* transcriptome assembly and annotation

The Ceratopteris transcriptome was assembled from approximately ~188 million paired end reads from three biological replicates of $-A_{CE}$ treated gametophyte cDNA libraries and ~207 million reads from three biological replicates of $+A_{CE}$ treated gametophyte cDNA libraries; Table 2.2 provides a summary of run metrics, analysis and assembly of the transcriptome. After removing adapter sequences and reads mapping to contaminants, the remaining reads were used to assemble a reference Ceratopteris transcriptome using Trinity (Grabherr et al., 2011b); 206,059 predicted transcripts (including isoforms) were assembled using a minimum length cutoff of 150. The distribution of the read depth across all putative genes is shown in Figure 2.2. Of the 111,977 putative, unique genes, 82,820 had read support; 38% of the read-support genes had BLASTx hits to the nr database (E-value $<1x10^{-10}$), while 34% could be mapped to GO terms using BLAST2GO (Aparicio et al., 2006; Conesa & Gotz, 2008; Conesa et al., 2005; Gotz et al., 2008). The GO terms associated with the entire Ceratopteris transcriptome is shown in Table 2.3.

The quality of the Trinity assembly was assessed by using BLASTn to compare the 5,133 Ceratopteris Sanger EST sequences available in GenBank to the transcript sequences generated by Trinity. 87% of the Sanger ESTs were identical or almost identical (E-value of 0.0) to transcripts in the transcriptome assembly, indicating that Trinity accurately assembled transcript sequences from the short Illiumina reads.

2.3.3 Identification of differentially expressed genes by A_{CE} treatment

Three programs were used to identify differentially expressed genes: edgeR (M. D. Robinson et al., 2010), DESeq (Anders & Huber, 2010) and EBSeq (Leng et al., 2013). With edgeR and DESeq, the False Discovery Rate was controlled at q=0.01 using the approach by Benjamini and Hochberg (Benjamini et al., 2001). With EBSeq, the posterior probability cutoff was set to 0.99. An additional practical significance cutoff of at least a two-fold difference in expression was also applied. A scatterplot (Fig. 2.3A) was used to assess the overall expression pattern across all transcripts in the transcriptome. As seen by its linear trend, the expression of the vast majority of transcripts was similar regardless of treatment, as expected. The majority (88%) of differentially expressed genes were more highly expressed in $+A_{CE}$ treated than $-A_{CE}$ treated gametophytes (Fig. 2.3B). The number of differentially expressed genes identified varied slightly depending upon the statistical model used (Fig.2.4). DESeq was the most conservative, identifying 1,183 genes as differentially expressed, EBSeq identified 3,065 genes as differentially expressed, and edgeR, the least conservative, identified 3,700 genes as differentially expressed. The 1,163 genes found to be differentially expressed by all three packages were used in subsequent analyses; their associated GO terms are shown in Table 2.3. Differences in gene expression were validated by qRT-PCR for 10 genes including genes up-regulated in $+A_{CE}$ samples, genes up-regulated in $-A_{CE}$ samples and genes showing no significant differences in expression. As shown in Figure 2.5, the qRT-PCR expression data agrees with the RNA-Seq expression data for eight of the ten genes. The trends of the RNA-Seq data and qRT-PCR data agree for ten out of ten genes.

2.3.4 Identification of candidate genes of the sex-determining pathway

The sex determination pathway in Ceratopteris, which is based upon the epistatic interactions among >70 sex-determining mutants (Banks, 1994b, 1997b, 1997d; Strain et al., 2001) is shown in Figure 2.6. In this model, there are two major regulatory genes that determine the sex of the gametophyte: the *TRANSFOMER* (*TRA*) and *FEMININIZATIOM* (*FEM*) genes. The *TRA* gene promotes the development of female traits (meristem and archegonia) because *tra* mutants are always male even in the absence of A_{CE}. The *FEM* gene is necessary for the development of male traits (antheridia) because the *fem* mutants are always female in the presence of A_{CE}. *TRA* and *FEM* also repress each other such that only one can be expressed (Banks, 1997d). The presence or absence of A_{CE} determines whether *TRA* or *FEM* is expressed: in the presence of A_{CE} *FEM* is expressed whereas in the absence of A_{CE} *TRA* is expressed. A_{CE} is perceived and transduced by the *HERMAPHRODITIC* (*HER*) genes; *her* mutants secrete A_{CE} but are A_{CE}-insensitive and develop as hermaphrodites in its presence. When A_{CE} is present, the *HER* genes act to repress *TRA*; because *TRA* represses *FEM*, *FEM* is expressed and the gametophyte develops as a male. When A_{CE} is absent, *TRA* is not repressed, *TRA* represses *FEM* and the gametophyte develops female traits. This pathway is remarkably similar to the GA signaling pathway in Arabidopsis as well as the recently described antheridiogen signaling pathway in the fern *Lygodium japonicum* (Tanaka et al., 2014), which also has an antheridiogen response. In Arabidopsis, GA binds to its receptor (*GID1*) and forms a complex with SCF^{SLY}/*GID2* that ultimately degrades the DELLA transcription factors responsible for repressing GA responses (reviewed in (Daviere & Achard, 2013; Sun, 2011). In *L. japonicum*, its antheridiogen binds to the GID receptor, which results in the

degradation of a *L. japonicum* DELLA protein in gametophytes (Tanaka et al., 2014). While the specific responses to GA in angiosperms and antheridiogens in fern gametophytes differ, the similarities of the pathways raise the possibilities that the *HER* genes are homologs of *GID1* or *SCR^{SLY}/GID2* and that *TRA* is a homolog of a DELLA–encoding gene. Genes very similar to *GID1, SCR^{SLY}/GID2* and *GAI*, a DELLA domain transcription factor, are present in the Ceratopteris transcriptome (alignments are shown in Fig. 2.7) but are not differentially expressed.

In Arabidopsis, the *GAMYB* transcription factor *MYB33*, originally identified as one of three homologs of the activator of GA-induced amylase expression in barley aleurone (Gubler, Chandler, White, Llewellyn, & Jacobsen, 2002; Gubler et al., 1995), is a core regulator of GA-induced responses (Gocal et al., 2001); it is a target of DELLA repression and is de-repressed in the presence of GA. Four genes with MYB domains are up-regulated by $+A_{CE}$ treatment in Ceratopteris (Table 2.4) and we predict that the *FEM* gene may encode one of these *MYB* genes. Support for this prediction comes from the recent characterization of two *GAMYB* genes (*PpGAMYB1* and *PpGAMYB2*) in *Physcomitrella patens,* which are also similar to *MYB33* and comp82703, one of the four MYB genes in Ceratopteris (Table 2.4). Knocking-out *PpGAMYB2* in Physcomitrella leads to gametophytes with fewer antheridia and more archegonia, suggesting that *PpGAMYB2* promotes the differentiation of sperm-forming antheridia and suppresses egg-forming archegonia formation in Physcomitrella (Aya et al., 2011), as does the *FEM* gene in Ceratopteris gametophytes (Strain et al., 2001).

Among the genes up-regulated by $-A_{CE}$-treatment is a gene similar to *COPALYL DIPHOSPHATE SYNTHASE/ENT-KAURENE SYNTHASE* (*CPS/KS*), which encodes a

key enzyme in GA biosynthesis (Hedden & Thomas, 2012; Sun & Kamiya, 1994). In *L. japonicum*, *CPS/KS* is also more highly expressed in gametophytes that secrete antheridiogen (Tanaka et al., 2014). As illustrated in Figure 2.6, we propose that the product of the *FEM* gene acts directly or indirectly to down-regulate *CPK/KS* expression, but only in males. The rationale for this interaction is based on the knowledge that A_{CE} is secreted by the hermaphrodite but not the male (Banks et al., 1993c). Because the *FEM* gene or gene product is repressed in the hermaphrodite, we predict that *CPK/KS* is a target of repression by *FEM* and is down-regulated in $+A_{CE}$ treated gametophytes rather than up-regulated in $-A_{CE}$-treated gametophytes. In other words, *FEM* prevents A_{CE} production in the male by down-regulating *CPK/KS* expression.

Whether any of the sex-determining genes in Ceratopteris are actually encoded by the genes described can be tested either by sequencing the relevant genes in the appropriate mutants and comparing them to the corresponding wild-type sequences, or by overexpressing or knocking-down the expression of candidate genes and examining their effects. Having a Ceratopteris transcriptome has and will be invaluable for these experiments to proceed.

2.3.5 Genes up-regulated in $-A_{CE}$ treated samples

Of the 133 genes that are up-regulated by $-A_{CE}$ treatment (or down-regulated in $+A_{CE}$ treated samples), 55% were annotated as protein-encoding genes (Table 2.4). In addition to the *CPS/K*S gene previously described, several genes involved in hormone biology were found to be up-regulated by $-A_{CE}$ treatment. They include genes similar to *ABA 8'HYDROXYLASE*, which is involved in ABA catabolism (Kushiro et al., 2004), the

transcription factors *ABF2/ABRE1* and *ARIA* involved in ABA regulated gene expression (Cutler, Rodriguez, Finkelstein, & Abrams, 2010; Fujita, Fujita, Shinozaki, & Yamaguchi-Shinozaki, 2011), two A-type response regulators that are involved in cytokinin-mediated signaling (W. Zhang, To, Cheng, Schaller, & Kieber, 2011), and *KUF1*, an F-box protein up-regulated by karrikins (S. M. Smith & Li, 2014). While ABA is known to affect sex determination by blocking the A_{CE} response (Hickok, 1983), these results indicate a role for other hormones in the sex-determining process.

Four putative cytochrome P450 monooxygenases are up-regulated in the -A_{CE} sample. While the functions of these genes are unknown, one is notable in that its expression is elevated 137-fold in the -A_{CE} samples (Table 2.4). In contrast to the genes that are up-regulated in the +A_{CE} samples, only two transposon sequences and no genes encoding protein kinases or proteins involved in chromatin modification or other epigenetic marks were found among the genes up-regulated in the -A_{CE} samples.

2.3.6 The response to A_{CE}- transposon activation, chromatin remodelin, and epigenetic reprogramming of the gametophyte

Of the 1030 genes that are expressed at least two-fold higher in +A_{CE} samples, 723 (71%) could be annotated by Blast2GO. The classes of protein-coding genes well represented in these samples (Tables 2.4) include those similar to genes involved in hormone biology (20 genes), transcription (26 genes), chromatin organization or remodeling (31 genes), small RNA biogenesis and function (8 genes), RNA splicing, polyadenylation, stability and decay (11 genes), and protein processing (11 genes), as well transposon related transcripts (41). By extrapolating from what is understood about

the functions of many of these genes in other plants, several reasonable and testable hypotheses emerge regarding the molecular mechanisms underlying the response to A_{CE} in Ceratopteris.

Almost all transposon-related transcripts were annotated as retroelements (particularly *Copia* and *Gypsy* LTR retrotransposons) and up-regulated between 2.5- and 14.6-fold in the $+A_{CE}$ samples. Their abundance in these samples indicates that transposons are actively transcribed in gametophytes destined to become male. In Arabidopsis mature pollen (the male gametophyte), transposons are transcribed in the vegetative nucleus but not the sperm nuclei. Small interfering RNAs (siRNAs) originating from transposons in the vegetative nuclei are transported into the sperm nuclei to further silence the transposons in the sperm (Martienssen & Chandler, 2013; Slotkin et al., 2009). Transposon reactivation following A_{CE} exposure may, therefore, serve to reinforce transposon silencing and limit transposon-mediated genome instability in cells destined to become sperm later in male gametophyte development.

A striking number of genes up-regulated by $+A_{CE}$ treatment encode proteins that are involved in transcriptional reprogramming of the genome (Table 2.4). They include genes similar to the DNA methylation genes *DNA METHYLTRANSFERASE 1* (*MET1*), which maintains CpG methylation (Jullien, Susaki, Yelagandula, Higashiyama, & Berger, 2012; Saze, Mittelsten Scheid, & Paszkowski, 2003), *CHROMOMETHYLASE 3* (*CMT3)*, which maintains CpHpG methylation (Law & Jacobsen, 2010) and *NEEDED FOR RDR2-INDEPENDENT DNA METHYLATION* (*NERD*), which is involved in methylation of transcriptionally silent regions (Pontier et al., 2012). Other genes similar to those involved in transcriptional silencing in Arabidopsis that are up-regulated include *DICER*-

LIKE3 (*DCL3*), which functions in RDR2 dependent small interfering RNA (siRNA) production (I. R. Henderson et al., 2006), the histone H3 lysine 9 (H3K9) methyltransferases *KRYPTONITE* (*KYP*) which is required for DNA methylation (Jackson, Lindroth, Cao, & Jacobsen, 2002) and the H3K9 methyltransferase *SUVH6* homologs (Ebbs & Bender, 2006). Genes similar to the second largest subunit of the plant specific DNA DEPENDENT RNA POLYMERASE IV and/or V (*NRPD2A* and *NRPD2B*), required for the production of siRNAs and for RdDM in Arabidopsis (Onodera et al., 2005) are also up-regulated by $+A_{CE}$ treatment. Interestingly, *ROS1* (*REPRESSOR OF SILENCING 1*), a DNA demethylase (Gong et al., 2002) is also up-regulated by A_{CE}, as is the histone acetyltransferase *INCREASED DNA METHYLATION 1* (*IDM1*) involved in DNA demethylation (Qian et al., 2012), which may contribute to reprogramming of DNA methylation leading to loss of silencing at some loci (Zhu, Kapoor, Sridhar, Agius, & Zhu, 2007). A gene similar to the Arabidopsis *METHYL-CYTOSINE BINDING DOMAIN 9* (*MBD9*) was also found to be up-regulated by A_{CE} (M. Peng, Cui, Bi, & Rothstein, 2006). While we were able to identify genes similar to other components of the gene and transposon silencing pathways (reviewed and listed in (Matzke & Mosher, 2014), their transcript abundance is unaffected by $+A_{CE}$ treatment (data not shown).

RdDM was not the only transcriptionally repressive process up-regulated by $+A_{CE}$ treatment, as we also identified a gene similar to the histone H3K27 methyltransferase CLF up-regulated by $+A_{CE}$ treatment 2.8 fold (Table 2.4). This leads to the hypothesis that Polycomb silencing via histone H3K27 methylation also plays a role in epigenetic reprogramming early in the establishment of the male developmental program. While the

targets of Polycomb silencing in the fern gametophytes remain to be discovered, our results point to a role for SWN in determinate growth of the male gametophyte, similarly to its role in the moss *Physcomitrella patens* (Okano et al., 2009). Active chromatin marks, particularly H3K4 di- and tri-methylation (H3K4me2 and H3K4me3), are conferred by the Trithorax class of histone methyltransferases (Schuettengruber, Chourrout, Vervoort, Leblanc, & Cavalli, 2007). $+A_{CE}$ treatment up-regulates a homolog of *ATXR3* (*SDG2*), a H3K4me3 methyltranferase required for gametophyte development in Arabidopsis (Berr et al., 2010) (Table 2.4), as well as a homolog of the H3K4me2 methyltransferase *ATX2* (*SDG30*), which has been shown to be expressed during Arabidopsis anther development (Saleh et al., 2008). The histone H3 lysine36 methylatransferase *EFS* (*SDG8*) homolog was also up-regulated (3-fold) by $+A_{CE}$ treatment (Table 2.4). Mutants of *EFS* (*SDG8*) have a pleiotropic effect on plant development in Arabidopsis, including pollen development (Grini et al., 2009).

Chromatin remodeling plays an integral role in the establishment of transcriptionally permissive chromatin states (Clapier & Cairns, 2009). Homologs of nine plant chromatin remodelers from the SWI/SNF family were up-regulated by $+A_{CE}$ treatment. These include two genes homologous to *PICKLE*, a positive regulator of GA response pathway (J. T. Henderson et al., 2004; Ogas, Kaufmann, Henderson, & Somerville, 1999), *BRAHMA (CHR2)* (Farrona, Hurtado, Bowman, & Reyes, 2004) and the chromatin remodeler genes *CHR11*, *CHR21/INO80* and *SPLAYED* (*CHR3*) all of which that have been implicated in gametophyte development and meristem maintenance in Arabidopsis (Huanca-Mamani, Garcia-Aguilar, Leon-Martinez, Grossniklaus, & Vielle-Calzada, 2005; Wagner & Meyerowitz, 2002).

The importance of chromatin and DNA modification-based epigenetic inheritance and imprinting, as well as transposon silencing during angiosperm gametophyte development, is well documented in plants (reviewed in (Borges, Calarco, & Martienssen, 2012). The observed differences in the expression of genes that are involved in chromatin and DNA modification in Ceratopteris suggest that sex determination by A_{CE} may involve extensive epigenetic reprogramming of the young male gametophyte genome. In Arabidopsis, a comparison of genome-wide DNA methylation patterns, small RNA populations and chromatin states of vegetative cells and their neighboring gametes reveals that extensive epigenetic reprogramming occurs during pollen and embryo sac development (Baroux, Raissig, & Grossniklaus, 2011; Borges et al., 2012; Calarco et al., 2012). Our results suggest that epigenetic reprogramming of the gametophyte may be a common feature of euphyllophyte gametophytes.

2.3.7 Hormone related genes up-regulated by $+A_{CE}$ treatment

Several cytokinin, auxin and ethylene related genes are up-regulated by $+A_{CE}$ treatment, including homologs of the cytokinin receptor genes *CYTOKININ-INDEPENDENT1* (*CKI1*) and *ARABIDOPSIS HISTIDINE KINASE 4* (*AHK4*) (Hwang, Sheen, & Muller, 2012), the auxin transport genes *BIG* (Gil et al., 2001), *ABCB19/PGP19/MDR1* (Noh, Murphy, & Spalding, 2001) and two *PIN-FORMED* (*PIN*) genes (Petrasek et al., 2006), and *ETHYLENE-INSENSITIVE PROTEIN 2* (*EIN2*), an activator of ethylene responses (Alonso, Hirayama, Roman, Nourizadeh, & Ecker, 1999). The up-regulation of these hormone related genes by $+A_{CE}$ treatment in Ceratopteris

suggests that A_{CE} may influence auxin, cytokinin and ethylene responses, or that the crosstalk among hormones modulates growth and differentiation of the male.

Several other transcription factor homologs up-regulated by $+A_{CE}$ treatment are associated with GA responses in angiosperms, including *MOTHER OF FT* (*MFT*) and three GRAS family transcription factors, including *SCARECROW* (*SCR*) and *LOST MERISTEMS* (*LOM*) (Table 2.4). Any of these transcription factors could be encoded by the FEM gene, or activated directly or indirectly by the FEM gene product. Of the remaining transcription factor homologs up-regulated by $+A_{CE}$ treatment (Table 2.4), several are known for their role in diverse developmental processes in Arabidopsis and include three *HD-Zip* genes. We speculate that these genes could affect patterns of cell division that distinguish males from hermaphrodites.

The final noteworthy class of genes up-regulated by $+A_{CE}$ treatment includes those involved in protein processing. Of these 11 genes, five are homologs of E3 ubiquitin ligases and four are ubiquitin related proteins (Table 2.4). In Arabidopsis, the GA (and other hormone) signaling pathway requires the degradation of ubiquitinated proteins, including the DELLA family of transcriptional repressors of GA responses (Santner & Estelle, 2010; Shabek & Zheng, 2014) via the 26S proteasome. The up-regulation of these genes by A_{CE} treatment lends further support to the possibility that $+A_{CE}$ signaling in Ceratopteris is similar to GA signaling in Arabidopsis at the molecular level.

2.3.8 Notes

SEM photos were taken with the help of the Purdue Microscopy Facility. This chapter was written for submission to a peer-reviewed journal with Michael Gribskov, Federico Gaiti, Olga Vitek, Milos Tanurdzic, and Jo Ann Banks.

2.3.9 Accession Numbers

The transcriptome shotgun assembly project has been deposited at DDBJ/EMBL/GenBank under the accession SAMN02821161.

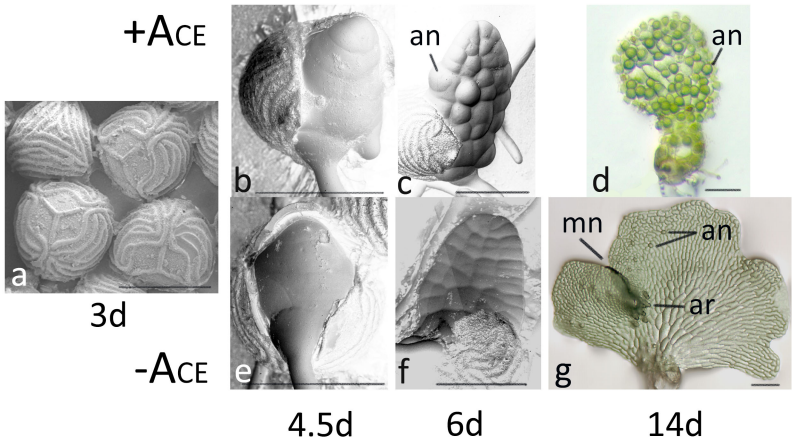


Figure 2.1. Gametophyte morphology. (**A**) SEM of spores three days after inoculation. The spores have yet to burst at their trilete markings. (**B**) and (**C**) SEMs of 4.5d and 6d old gametophytes grown in the presence of ACE. (**D**) A 14d old mature male showing numerous antheridia (an). (**E**) and (**F**) SEMs of 4.5d and 6d old gametophytes grown in the absence of ACE. (**G**) A 14d old mature hermaphrodite with a meristem notch (mn), archegonia (ar) and antheridia (an). Bars = 100mM.

Table 2.1. Primers used for qRT-PCR.

Gene	**Forward Sequence**	**Reverse Sequence**
CrEF1α	5'CAGACCAGTCGGAGCAAAAGT	5'TCCTGTGGGAAGGGTGGAA3'
comp39080	5'CGCAAGGGATAGCCAAATTA3'	5'CGATCTCAACGCGATCTACA3'
comp82638	5'CTGCTGCCTCTCAGTGTGAC3'	5'ATCACGCGCTTGTAGGACTT3'
comp114251	5'AGCTCAAATGCCACCACTTT3'	5'ACATAGCCGCTGCTGTTCTT3'
comp38095	5'ATGCCGAATGGAAGACTGTT3'	5'TTCATATTCGGCGACTCCTT3'
comp82048	5'GGTATGACGCCACAGAACCT3'	5'TGCAGACATTGCAGGATACC3'
comp103387	5'TCGAAAGAGAGGCAACACCT3'	5'ACTTTCCGAGAAGCAGTGGA3'
comp46913	5'TGGGCAAACTTCAGGTAAGG3'	5'TGAGGCTGTGTCAGAGATGC3'
comp105977	5'AGGAAATCGCTGGACGTAGA3'	5'CCTCATCCTTCCAACATCGT3'
comp110703	5'GAGGTAAGGCAAGCGCTCTA3'	5'CCAACGGCCATGAGAAGTAT3'
comp109704	5'GGCGAAATACCTGCAAATGT3'	5'TCACGACACACAACCACAGA3'
comp84184	5'ATGGGCAGATGGTGGAAATA3'	5'TGACCATTGTCTCCCTCAGA3'

Table 2.2. Run metrics, assembly and analysis statistics for the combined, $-A_{CE}$ and $+A_{CE}$ treatment datasets.

	Combined Data Set	$-A_{CE}$	$+A_{CE}$
Run Metrics			
Total bases	39,944,451,822	19,004,923,762	20,939,528,060
Total reads	395,489,622	188,167,562	207,322,060
Average GC%	46.88	47.40	46.35
% with Phred scores >20	90.53	88.90	92.15
% with Phred scores >30	81.33	78.21	84.45
Contaminant reads removed	98,989,731 (25%)	86,943,515 (46%)	12,046,216 (6%)
hits to bacteria	2,233,971	1,650,498	583,473
hits to viruses	1,160,904	998,639	162,265
hits to rRNA	98,654,852	87,216,917	11,436,935
hits to chloroplast	6,897,428	5,854,557	1,042,871
hits to mitochondria	6,681,340	5,599,946	1,081,394
total contaminant hits	115,628,495	101,320,557	14,306,938
Analysis			
DESeq DEGs	1183	140	1043
edgeR DEGs	3700	1585	2115
EBSeq DEGs	3065	1065	2000
Intersection of DEGs	1163	133	1030
Assembly			
Total transcripts assembled	206,059		
Total genes assembled	111,977		
N50	1,988		
Min length	151		
Max length	17,306		
Average length	867		
% Reads aligned in RSEM	87.7		
Genes with read support	82,820		

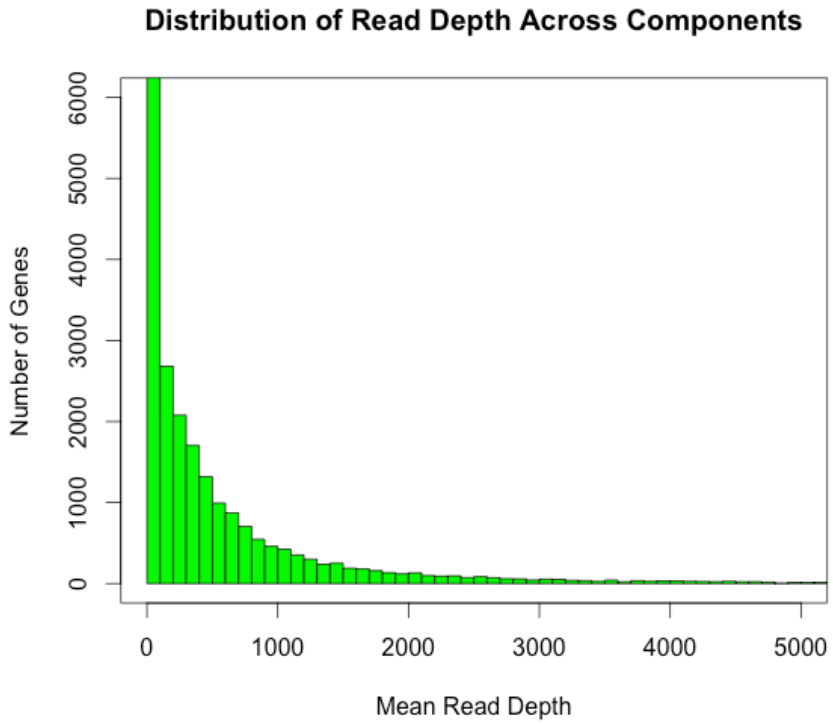


Figure 2.2. Histogram depicting the distribution of normalized read count across the 82,820 components that had at least one read align across all samples.

Table 2.3 Table of GO terms. GO terms mapping to the whole assembly and to DEGs

GO term	% of sequences with GO term
Biological process GO terms for all transcripts with read support	
cellular developmental process	3
transmembrane transport	6
small molecule biosynthetic process	5
single-organism carbohydrate metabolic process	3
signal transduction	6
response to oxygen-containing compound	3
response to inorganic substance	4
response to hormone stimulus	5
response to abiotic stimulus	3
reproductive structure development	3
regulation of transcription, DNA-dependent	4
regulation of biological quality	3
protein phosphorylation	3
post-embryonic development	3
oxidation-reduction process	3
organonitrogen compound biosynthetic process	3
organic substance transport	3
organic substance catabolic process	3
DNA metabolic process	5
RNA processing	3
anatomical structure morphogenesis	3
carbohydrate derivative metabolic process	3
carboxylic acid metabolic process	4
cell cycle	4
cellular catabolic process	3
cellular component biogenesis	5
Cellular component GO terms for all transcripts with read support	
integral to membrane	8
vacuolar membrane	3
ribosome	3

Table 2.3 Continued

protein complex	11
plastid thylakoid membrane	2
plasmodesmata	4
plasma membrane	16
nucleolus	3
mitochondrial part	2
microtubule cytoskeleton	3
Golgi apparatus	5
cell wall	3
chloroplast envelope	4
chloroplast stroma	4
chloroplast thylakoid	3
cytoplasmic membrane-bounded vesicle	3
cytoskeletal part	3
cytosol	11
endoplasmic reticulum	3
endosome	2
extracellular region	4
Molecular function GO terms for all transcripts with read support	
inorganic cation transmembrane transporter activity	3
isomerase activity	3
ligase activity	3
zinc ion binding	5
transferase activity, transferring one-carbon groups	2
transferase activity, transferring hexosyl groups	2
structural constituent of ribosome	2
signal transducer activity	3
sequence-specific DNA binding transcription factor activity	4
protein serine/threonine kinase activity	7
protein dimerization activity	3
phosphatase activity	3
peptidase activity, acting on L-amino acid peptides	3
oxidoreductase activity	13
nucleotidyltranferase activity	2
lyase activity	3
ATP binding	16
ATPase activity, coupled	4
DNA binding	10

Table 2.3 Continued

hydrolysis-driven transmembrane transporter activity	2
hydrolase activity, acting on glycosyl bonds	3
Biological process GO terms for DEGs	
cellular protein modification process	7
regulation of gene expression, epigenetic	3
phyllome development	3
root development	3
response to other organism	3
post-embryonic organ development	3
response to inorganic substance	3
regulation of developmental process	3
positive regulation of cellular process	3
epidermal cell differentiation	3
single-organism carbohydrate metabolic process	4
carbohydrate derivative metabolic process	4
cell development	4
signal transduction	4
cellular component biogenesis	4
cell cycle process	4
phosphorylation	4
regulation of biological quality	4
flower development	4
response to hormone stimulus	4
organonitrogen compound metabolic process	5
DNA metabolic process	5
response to oxygen-containing compound	5
regulation of transcription, DNA-dependent	5
single-organism transport	5
response to oxygen-containing compound	5
regulation of transcription, DNA-dependent	5
single-organism transport	5
Sequence distribution of cellular component GO terms for DEGs	
cytosol	13
endomembrane system	2
vacuolar membrane	3
ribonucleoprotein complex	3
plasmodesmata	13
plasma membrane part	2

Table 2.3 Continued

plant-type vacuole	2
organelle inner membrane	2
nucleoplasm	2
nucleolus	2
mitochondrial membrane	2
microtubule	3
integral to membrane	10
endosome	2
endoplasmic reticulum	2
Golgi apparatus	6
apoplast	3
cell wall	5
chloroplast envelope	5
chloroplast stroma	6
chloroplast thylakoid membrane	3
chromosome	2
cytoplasmic membrane-bounded vesicle	6
Sequence distribution of molecular function GO terms for DEGs	
ATP binding	25
hydrolase activity, acting on glycosyl bonds	3
cation-transporting ATPase activity	3
metal-ion transporter	3
signaling receptor	3
methyltransferase	3
structural molecule activity	3
nucleotidyltransferase	3
microtubule motor activity	3
transcription factor	6
protein serine/threonine kinase	6
zinc ion binding	7

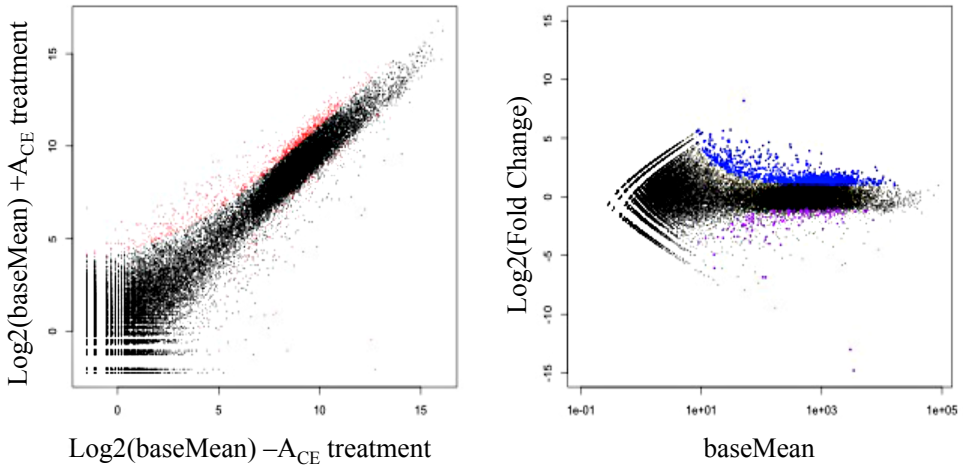


Figure 2.3. Visual representation of the differentially expressed genes. These plots show the 1163 genes were found to be differentially expressed at a 0.01 FDR with at least a 2 fold change. A. The expression scatterplot shows the $\log_2$(baseMean) (the base mean is the counts corrected for library size differences)for the hermaphrodite gametophytes (-A_{CE}) vs. The $\log_2$(baseMean) for the male gametophytes (+A_{CE}). The genes that are differentially expressed are shown in red. The plot shows a linear trend, indicating that the majority of genes are equivalently expressed between samples. B. An MA plot showing the baseMean (in this plot the counts were corrected for differences in library conditions and then averaged across conditions) versus the $\log_2$(FoldChange). Genes that are up-regulated in males are blue, genes up–regulated in hermaphrodites are purple. In both plots, it is clear that the majority of the differentially expressed genes are more highly expressed in the male samples than in the hermaphrodite samples.

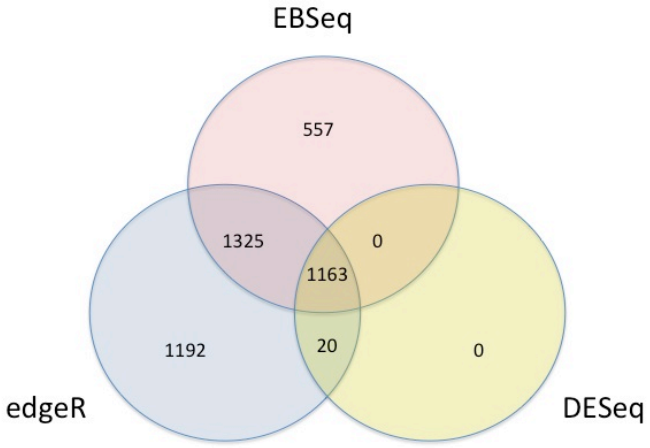


Figure 2.4. Venn diagram of genes called as differentially expressed in each of the three employed Bioconductor programs.

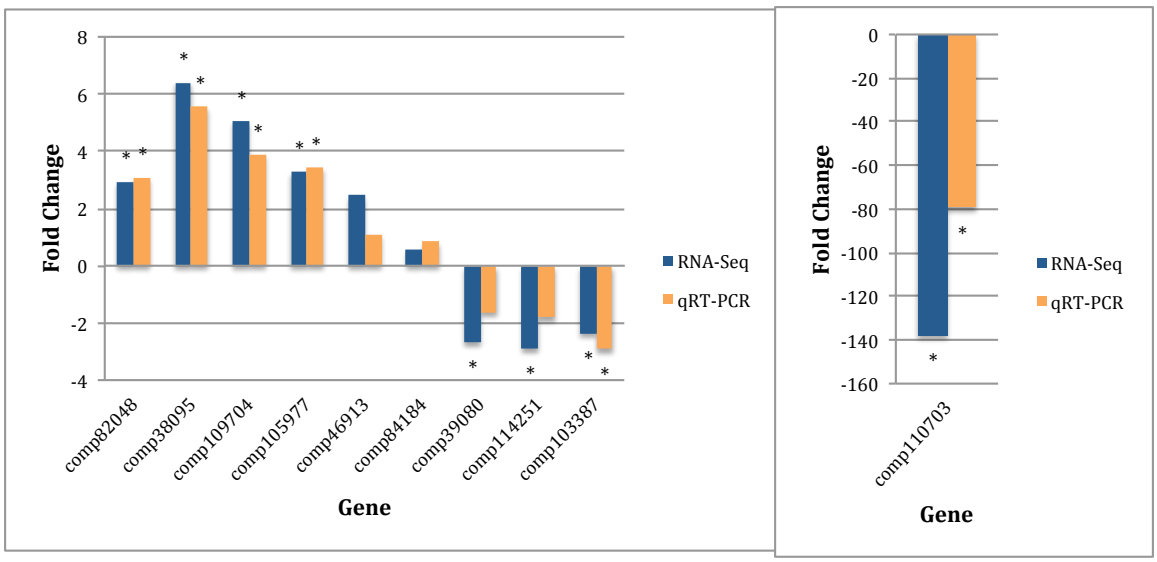


Figure 2.5. Comparison of gene expression from qRT-PCR vs. RNA-Seq. The fold changes for the qRT-PCR data were calculated using the ΔCt method (Livak & Schmittgen, 2001). A positive fold change value indicates that the gene was more highly expressed in $+A_{CE}$ samples, a negative fold change value indicates that the gene is more highly expressed in $-A_{CE}$ samples; *indicates that fold changes in expression are statistically significant.

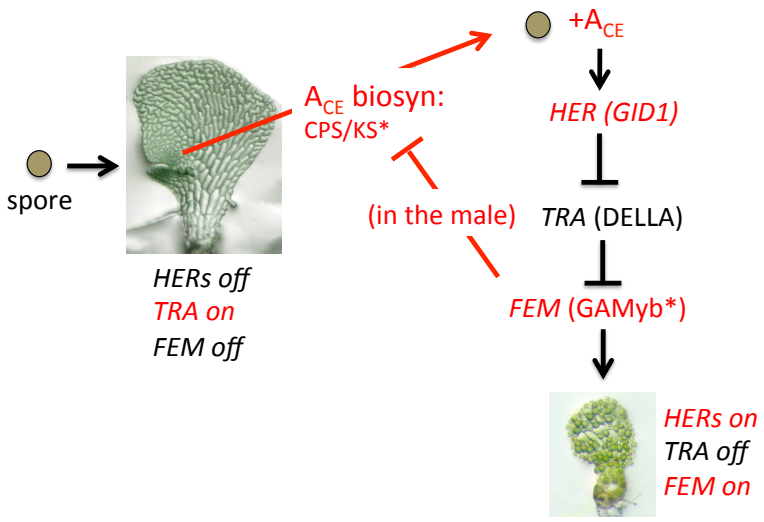


Figure 2.6. A model of the sex-determining pathway in *Ceratopteris*. The interactions among the *HER, TRA* and *FEM* genes are based on the epistatic interaction among these genes. Lines ending in arrows indicate positive interactions and lines ending in bars indicate repressing interactions. The candidate genes encoded by *HER, TRA* and *FEM* are shown in parenthesis. *FEM* is shown to prevent A_{CE} synthesis in the male by repressing *CPS/KS*, a key enzyme in A_{CE} biosynthesis.

Table 2.4. List of *Ceratopteris* genes mentioned in the discussion that are differentially expressed by A_{CE} treatment and are homologous to *Arabidopsis* genes.

Ceratopteris **gene number**	***Arabidopsis*** **homolog**	***Arabidopsis*** **Accession**	**BLASTx E-value**	**Fold Change**	**Adj Pvalue**
Genes upregulated by -A_{CE} treatment					
Hormones					
comp103387[a]	ABA 8'-hydroxylase	AT4G19230.1	0	2.4	3.23E-03
comp80125	ARR9	AT2G41310.1	3.00E-42	5.3	1.18E-08
comp82535	ARR9	AT2G41310.1	2.00E-48	4.2	1.30E-08
comp119738	KAR-UP F-box 1	AT1G31350.1	4.00E-32	2.3	9.17E-05
Transcription factors					
comp112296	CPS/GA1	AT4G02780.1	3.00E-159	2.2	1.46E-03
comp106738	ERF/AP2 family	AT5G67190.1	1.00E-19	3.3	6.65E-05
comp83407	ERF/AP2 family	AT5G11590.1	1.00E-28	11.8	5.84E-05
comp106310	A20/AN1-like zinc finger family	AT1G12440.2	1.00E-23	3.4	9.04E-05
comp101713	A20/AN1-like zinc finger family	AT2G36320.1	5.00E-31	2.2	5.09E-03
Secondary metabolism					
comp110703[a]	CYP76C2	AT2G45560.1	7.00E-87	125	5.88E-29
comp84540[a]	CYP75B1	AT5G07990.1	2.00E-101	6.8	7.19E-06
comp112472[a]	P450	AT3G26210.1	8.00E-53	3.3	1.44E-04
comp106199	CHS	AT5G13930.1	8.00E-116	2.4	9.66E-03
Genes upregulated by +A_{CE} Treatment					
GA					
comp116986	SCL	AT5G66770.1	1.00E-87	2.4	6.15E-03
comp82755 [a]	GRAS family	AT1G63100.1	1.00E-92	2.7	4.54E-04
comp103126	LOM	AT3G60630.1	5.00E-49	2.9	2.53E-05
comp81241	LRP	AT3G51060.1	2.00E-30	3.6	4.22E-06
comp42166	MFT	AT1G18100.1	5.00E-62	2.5	3.72E-04
ABA					
comp82182	ARM repeat protein	AT5G19330.1	0	2.2	5.71E-03
comp100365	ABI3	AT3G24650.1	2.00E-40	2.6	1.81E-04

Table 2.4 Continued

comp103619	PP2C	AT1G72770.3	1.00E-38	3.2	2.88E-05
comp114719[a]	KEG	AT5G13530.1	0	3.7	1.01E-07
Ethylene					
comp106297	EIN2	AT5G03280.1	1.00E-64	2.5	1.39E-03
Auxin					
comp101920[a]	NOV	AT4G13750.1	0	2.7	2.54E-04
comp106375	PIN4	AT2G01420.1	4.00E-166	4.6	2.68E-08
comp105872	PIN3	AT1G70940.1	5.00E-156	2.2	9.23E-03
comp98976	BIG	AT3G02260.1	0	4.2	7.20E-09
comp109704	ABC transporter	AT3G28860.1	0	4.7	7.48E-12
comp97116	DOT2	AT5G16780.1	3.00E-132	2.5	8.33E-03
comp114948	SAR1	AT1G33410.2	0	3	4.83E-05
comp105798	ARF	AT1G19220.1	5.00E-53	6.5	1.76E-04
Cytokinins					
comp111805	AHK4	AT2G01830.1	0	3.2	1.33E-04
comp100079	CKI1	AT2G47430.1	4.00E-108	2.6	7.26E-04
DNA methylation/demethylation					
comp115365	MET1	AT5G49160.1	0	3.3	1.45E-06
comp82159	CMT3	AT1G69770.1	1.00E-155	2.3	6.31R-03
comp112176[a]	ROS1	AT2G36490.1	8.00E-83	2.7	1.46E-03
comp101924[a]	NERD	AT2G16485.1	7.00E-96	3	4.01E-05
Chromatin remodeling					
comp109662	CHR11	AT3G06400.2	0	2.2	7.22E-03
comp83245[a]	CHR5	AT2G13370.1	0	3.9	2.28E-08
comp103550	CHR4	AT5G44800.1	0.00E+00	4.1	6.59E-09
comp40502	PKL	AT2G25170.1	0.00E+00	2.6	5.93E-04
comp103233	PKL/CHD3/CHR6	AT2G25170.1	5.00E-124	2.8	6.59E-05
comp39118	BRM	AT2G46020.2	0	5	5.18E-12
comp43532	CHR21/INO80			3	1.75E-05
Histone modification					
comp81987	MBD9	AT3G01460.1	5.00E-103	4.1	4.26E-09
comp99654	SUVH4/KYP	AT5G13960.1	0	2.5	1.19E-03
comp83034	CLF	AT2G23380.1	0	2.8	1.59E-04
comp102724	ATX2	AT1G05830.2	0	2.8	2.32E-04
comp83655	ATXR3	AT4G15180.1	2.00E-180	3.8	8.34E-08
comp98691	HAC12	AT1G16710.1	0	2.6	5.76E-04

Table 2.4 Continued

comp62161	HAC1	AT1G79000.1	0	2.6	1.02E-03
comp108638	HAC1	AT1G79000.1	0	2.5	3.35E-03
comp98650	Elongator subunit	AT5G13680.1	0	2.2	9.77E-03
comp106634[a]	EFS/SDG8	AT1G77300.2	2.00E-94	3.1	5.83E-06
comp110316[a]	IDM1	AT3G14980.1	1.00E-111	3	7.37E-05
comp111521	HDA14	AT4G33470.1	0	2.3	7.41E-03
comp109495[a]	SUVH6	AT2G22740.1	2.00E-142	2.5	7.20E-03
Other possible chromatin-related genes					
comp109512	RCC1	AT3G55580.1	4.00E-31	3	7.97E-05
comp37548	RCC1	AT5G19420.1	0	3	8.72E-05
comp103127[a]	FCA	AT4G16280.3	2.00E-67	3	9.77E-05
comp114220	ICU2	AT5G67100.1	0	2.9	4.69E-05
comp103536	Related to yeast Spt6 protein	AT1G65440.3	1.00E-94	3.4	3.47E-05
comp87951	TSO1	AT3G22780.1	1.00E-59	2.7	4.27E-04
comp102301[a]	EMB1691	AT4G09980.1	3.00E-150	2.5	1.32R-03
RNA processing					
comp100728	AtCSF77	AT1G17760.1	0	2.36	3.840E-03
comp81881	PCFS4	AT4G04885.1	1.00E-40	2.47	1.318E-03
comp81990	THO2	AT1G24706.2	0	3.01	2.46E-05
comp110109	PRP2	AT1G32490.2	0	2.67	2.838E-04
comp99888	splicing factor	AT1G60200.1	5.00E-67	2.64	3.316E-04
comp40366	splicing factor	AT1G80070.1	0	2.63	3.277E-04
comp102040[a]	mRNA splicing	AT3G52250.1	2.00E-25	3.54	6.15E-07
comp103037[a]	SUA	AT3G54230.2	2.00E-122	2.63	3.075E-04
comp106155[a]	SUA	AT3G54230.2	2.00E-73	2.9	1.04E-04
comp114187	UPF1	AT5G47010.1	0	2.74	1.233E-04
comp82523	CPSF160	AT5G51660.1	0	2.46	1.736E-06
small RNA-related					
comp108491	AGO1	AT1G48410.1	0	2.5	8.92E-04
comp82278	AGO1	AT1G48410.1	0	2.6	3.45E-04
comp112142	DCL1	AT1G01040.1	0	2.5	1.16E-03
comp110523	DCL1	AT1G01040.1	0	2.9	1.63E-03
comp37939	DCL4	AT5G20320.1	2.00E-179	2.4	4.11E-03

Table 2.4 Continued

comp82821	SUO	AT3G48050.2	3.00E-91	3.9	2.98E-08
comp81850	NRPD2a	AT3G23780.1	0.00E+00	2.2	8.85E-03
comp111720	NRPD2b	AT3G18090.1	0.00E+00	2.5	1.06E-03
Transcription factors					
comp81559[a]	F2K11.14 with jumonji domain	AT1G63490.1	0	3.3	2.14E-06
comp39222	NAM	AT5G04410.1	2.00E-73	2.4	1.18E-03
comp100922[a]	LHW	AT2G27230.2	4.00E-61	2.5	3.18E-03
comp97820	AtNLP9	AT3G59580.2	7.00E-121	2.5	1.73E-03
comp60977	CCR4-NOT transcription complex subunit 1	AT1G02080.2	0	2.5	1.02E-03
comp106858	WRKY42	AT4G04450.1	8.00E-54	2.6	4.74E-03
comp87951	TSO1	AT3G22780.1	1.00E-59	2.7	4.27E-04
comp81059	Squamosa promoter-binding protein-like	AT1G76580.1	2.00E-68	2.7	2.05E-04
comp81373	CCT/CRP/MED12	AT4G00450.1	0	2.9	9.00E-05
comp100517	GTA2	AT4G08350.1	0	3	1.26E-05
comp44064	SPT6-like protein	AT1G65440.2	0	3.4	7.31E-07
comp101812	HUA2	AT5G23150.1	2.00E-71	3.2	3.78E-06
comp40501	EDM2	AT5G55390.2	9.00E-101	2.2	7.17E-03
comp82703	MYB120/33/101 related	AT5G06100.2	4.00E-51	3.1	1.22E-05
comp82703	Physcomitrella GAMYB1		1.00E-58		
comp82703	Physcomitrella GAMYB2		3.00E-59		
comp91285	MYB	AT4G21440.1	6.00E-39	Inf	1.22E-05
comp106205	GAMYB/MYB101	AT2G32460.1	3.00E-41	Inf	3.75E-05
comp102904	MYB3R3	AT3G09370.1	7.00E-82	3.1	2.75E-05
comp99051	ALY3	AT3G21430.2	1.00E-119	3.1	1.07E-05
comp103183	RLT2	AT5G44180.1	0	3.4	2.53E-06
comp102650	RLT2	AT5G44180.1	0	3.7	2.20E-07
comp110663	PDF2	AT4G04890.1	0	2.9	8.01E-04

Table 2.4 Continued

comp105977	HDG2	AT1G05230.4	0	2.7	1.48E-04
comp42959	REV	AT5G60690.1	4E-41	4.3	1.17E-04
Protein processing					
comp103576	UFO	AT1G30950.1	3.00E-115	2.4	9.52E-03
comp40395	Ubiquitin carboxyl-terminal hydrolase-related	AT3G47890.1	0	2.4	2.45E-03
comp106922	C3HC4-type RING finger	AT5G60710.1	2.00E-132	2	3.02E-03
comp82087	ubiquitin protease.	AT5G06600.3	3.00E-36	3.1	1.10E-05
comp82979	DCAF/DWD protein	AT4G31160.1	0	3.2	5.71E-06
comp113654	HECT ubiquitin ligase	AT4G38600.1	0	3.3	1.11E-06
comp103433	E3 ubiquitin ligase	AT5G05560.1	0	3.4	4.25E-06
comp40443	SNF2 domain protein	AT3G54460.1	0	3	6.48E-05
comp115766	E3 ubiquitin ligase	AT5G22000.2	4.00E-31	2.6	9.91E-05
comp41292[a]	RING E3 ubiquitin ligase	AT2G22010.1	0	2.6	2.14E-04
comp110105	ATL6	AT3G05200.1	5.00E-29	3.9	1.07E-08

[a]Putative homology of the *Ceratopteris* gene to an *Arabidopsis* gene based only upon BLAST results, including reciprocal best blast hit.

```
Multiple sequence alignment of CPS/KS by MUSCLE (3.8)
KS-Arabidopsis          ------------------------------------------------------------
KS-rice                 ------------------------------------------------------------
comp112296_c0_seq1      MSCSGNMYIHCCYLPVCQIDMPIATCSTKRVTFQLLNGSSAIVLVRGRTNKCGTVLQCTL
CPS-Arabidopsis         --------MSLQYHVLNSIPSTTFLSSTKTTISSSFLTISGSPLNVARDKSRSGSIHCSK
CPS-rice                ------------------------------------------------------------


KS-Arabidopsis          ----MSINLRSSGCSSPISATLERGLDSEVQTRANNV----------------------
KS-rice                 ----------------------------MQHR----------------------------
comp112296_c0_seq1      KGSFRYACMPSTTACHVRLDTIAASLGELQRSSKPKEFSHGETDVPATMWLLQSTETQIS
CPS-Arabidopsis         LRTQEYINSQEVQHDLPLIHEW----QQLQGEDAPQI----------------------
CPS-rice                -------QANIIEHETPRITKWPNESRDLDDHQQNNE----------------------


KS-Arabidopsis          --------SFEQTKEKIRKMLEKV---ELSVSAYDTSWVAMVPSPSSQNAPLFPQCVKWL
KS-rice                 ----------KELQARTRDQLQTL---ELSTSLYDTAWVAMVPLRGSRQHPCFPQCVEWI
comp112296_c0_seq1      TAHANENEQIQHLILRVKAMFQNMNLGEVSLSSYDTAWVALVPSLHDPRIPQFPQCLDWI
CPS-Arabidopsis         -SVGSNSNAFKEAVKSVKTILRNLTDGEITISAYDTAWVALIDA--GDKTPAFPSAVKWI
CPS-rice                -ADEEADDELQPLVEQVRSMLSSMEDGAITASAYDTAWVALVPRLDGEGGTQFPAAVRWI
                                  :      .  : .:    :: * ***:***::    .   . ** .: *:

KS-Arabidopsis          LDNQHEDGSWGLDNHDHQSLKKDVLSSTLASILALKKWGIGERQINKGLQFIELNS-ALV
KS-rice                 LQNQQDDGSWG-TRGFGVAVTRDVLSSTLACVLALKRWNVGQEHIRRGLDFIGRNF-SIA
comp112296_c0_seq1      ERNQLPDGSWG-DKEMFLAFER--VCNTLACVVALKTWNRCRWGVQKGIDFIHRNIERMG
CPS-Arabidopsis         AENQLSDGSWG-DAYLFSYHDR--LINTLACVVALRSWNLFPHQCNKGITFFRENIGKLE
CPS-rice                VGSQLADGSWG-DEALFSAYDR--VINTLACVVALTRWSLHHDQCKQGLQFLNLNLWRLA
                          .*  *****          .  : .***.::**  *.      ..*: *:  *   :

KS-Arabidopsis          TDETIQKPTGFDIIFPGMIKYARDLNLTIPLGSEVVDDMIRKRDLDLKCDSEKFSKGREA
KS-rice                 MDEQIAAPVGFNITFPGMLSLAMGMDLEFPVRQTDVDRLLHLREIELEREAGDHSYGRKA
comp112296_c0_seq1      NEDEEYMPTAFEVVFPSLLEDARLLGLDLPYDSSVIQKLKRERERKKLEKIPLELVHKYPT
CPS-Arabidopsis         DENDEHMPIGFEVAFPSLLEIARGINIDVPYDSPVLKDIYAKKELKLTRIPKEIMHKIPT
CPS-rice                EEEPDTMPIGFEIAFPSLVEAARGLGIDFPYDHPALKGIYANRELKLKRIPKDMMHIVPT
                         ::    * .*:: **.::. *  :.: .*     :. :   .: .*   . .      :

KS-Arabidopsis          YLAYVLEGTRNLKDWDLIVKYQRKNGSLFDSPATTAAAFTQFGNDGCLRYLCSLLQKFEA
KS-rice                 YMAYVTEGLGNLLEWDEIMMFQRKNGSFFNCPSTTAATLVNHYNDKALQYLNCLVSKFGS
comp112296_c0_seq1      TLLHSLEGIHRLLDWDKILKLQTKNGSFLFSTASTACALKYTHDKRCLDYLNHVLEKFDE
CPS-Arabidopsis         TLLHSLEGMRDL-DWEKLLKLQSQDGSFLFSPSSTAFAFMQTRDSNCLEYLRNAVKRFNG
CPS-rice                SILHSLEGMPGL-DWQRLLKLQCSDGSFLFSPSATAYALMQTGDKKCFAYIDRIIKKFDG
                         : :  **   * :*: ::  * .:**:: ..::** ::    :. .: *:   :..*

KS-Arabidopsis          AVPSVYPFDQYARLSIIVTLESLGIDRDFKTEIKSILDETYRYWLRGDEEIC-------L
KS-rice                 AVPTVYPLNIYCQLSWVDALEKMGISQYFVSEIKSILDTTYVSWLERDEEIM-------L
comp112296_c0_seq1      AVPSVYPLDLFERLWMVDRLERLGISRYFGKEIKDALDYVYRCW--TDKGIAWAKDSNVL
CPS-Arabidopsis         GVPNVFPVDLFEHIWIVDRLQRLGISRYFEEEIKECLDYVHRYW--TDNGICWARCSHVQ
CPS-rice                GVPNVYPVDLFEHIWVVDRLERLGISRYFQREIEQNMDYVNRHW--TEDGICWARNSNVK
                        .**.*:*.: : .:  :  *: :**.. *  **:. :* .   *   :. *

KS-Arabidopsis          DLATCALAFRLLLAHGYDVSYDPLKPFAEESGFSDTLEGYVKNTFSVLELFKAAQ-S-YP
KS-rice                 DITTCAMAFRLLRMNGYHVSSVELSPVAEASSFRESLQGYLNDKKSLIELYKASKVSKSE
comp112296_c0_seq1      DADDTAMAFRILRLHGYPVSPEVFYRFKKDGQFYCFEGETRQSVTGMFNLNRAAQIQ-FP
CPS-Arabidopsis         DIDDTAMAFRLLRQHGYQVSADVFKNFEKEGEFFCFVGQSNQAVTGMFNLYRASQLA-FP
CPS-rice                EVDDTAMAFRLLRLHGYNVSPSVFKNFEKDGEFFCFVGQSTQAVTGMYNLNRASQIS-FP
                        :    *:***:*  :** **   :  . : . *        :   .: :* .*::

KS-Arabidopsis          HESALKKQCCWTKQYLEM--ELSSWVKTSVRDKYLKKEVEDALAFPSYASLERSDHRRKI
KS-rice                 NESILDSIGSWSGSLLKE-----SVSSNGVKKAPIFEEMKYALKFPFYTTLDRLDHKRNI
comp112296_c0_seq1      DERILEEVFTFTESFLKQRRSLGRMKDKWVMSRGIREEVSYTLEFPWWKSLQRVEARQYI
CPS-Arabidopsis         REEILKNAKEFSYNYLLEKREREELIDKWIIMKDLPGEIGFALEIPWYASLPRVETRFYI
CPS-rice                GEDILQRARNFSYEFLRREREAQGTLHDKWIISKDLPGEVQYTLDFPWYASLPRVEARTYI
                         *  *.    :: . *          .. :    :  *:  :* :* : :* * : .  *

KS-Arabidopsis          LNGSAVENTRVTKTSYRLHNICTSDILKLAVDDFNFCQSIHREEMERLDRWIVENRLQEL
KS-rice                 ERF-DAKDSQMLKTEYLLPH-ANQDILALAVEDFSSSQSIYQDELNYLECWVKDEKLDQL
comp112296_c0_seq1      KHY-NVDDAWIAKSLYRMPFINNEVFRSLAILDYNKCQSIHQKELSKVLMWNQQSGFDKL
CPS-Arabidopsis         DQYGGENDVWIGKTLYRMPYVNNNGYLELAKQDYNNCQAQHQLEWDIFQKWYEENRLSEW
CPS-rice                GQYGGNDDVWIGKTLYRMPIVNNATYLELAKQDFNRCQALHQHELQGLQKWFIENGLEAF
                         .    .: .: *: * :    .     **  *:. .*: :. * . .  *  :. :.
```

```
KS-Arabidopsis          KFARQKLAYCYFSGAATLFSPELSDARISWAKGGVLTTVVD-DFFDVGGSK----EELEN
KS-rice                 PFARQKLTYCYLSAAATIFPRELSEARIAWAKNGVLTTVVD-DFFDLGGSK----EELEN
comp112296_c0_seq1      SFARQKPTECFFSIAATLFEPEFAYARIVWTQISVLVTLID-DLYDVKGSP----VDLER
CPS-Arabidopsis         GVRRSELLECYYLAAATIFESERSHERMVWAKSSVLVKAISSSFGESSDSRRSFSDQFHE
CPS-rice                GMTPEDVLRAYFLAAACIFEPNRASERLAWARVSVLANTISRHFYSDMSSM----KRMER
                         .  ..   .:   ** :*  : :  *: *:. .**.. :.  : .  .*       :

KS-Arabidopsis          LIHLVEKWDLN------GVPEYSSEHVEIIFSVLRDTILETGDKAFTYQG--RNVTHHIV
KS-rice                 LIALVEKWDGH------QEEFY-SEQVRIVFSAIYTTVNQLGAKASALQG--RDVTKHLT
comp112296_c0_seq1      FINALKRWDPK------EVETL-SEDTKIVYNGLYNTINMIGKETIACQD--RDFTLYIR
CPS-Arabidopsis         YIANARRSDHHFNDRNMRLDRPGSVQASRLAGVLIGTLNQMSFDLFMSHG--RDVNNLLY
CPS-rice                FM-----WSSLYEENGNVLGLEGYAKDGILARTLCQLIDLLSQETPPVREGQKCIHNLIR
                         :      .                    :   :   :   . .    .   . .   :

KS-Arabidopsis          KIWLDLL--KSMLREAEWSSDKSTPSLEDYMENAYISFALGPIVLPATYLIGPPLPEKTV
KS-rice                 EIWLCLM--RSMMTEAEWQRTKYVPTMEEYMANAVVSFALGPIVLPTLYFVGPKLQEDVV
comp112296_c0_seq1      ELVERFV--DSMHMESKWKAHQSFPTLEEYMENGKASIAVEAIIQISSFFLGEKILEEWF
CPS-Arabidopsis         LSWGDWM--------EKWKL---------YGDEGEGELMVKMIILMK--------NNDLT
CPS-rice                CAWIEWMMQQINMKDGRYDKGRVMHPGSCTVHNKETCLLIAQIVEICAGRIEE--AASMI
                              :          :.             :    : :  *:             .

KS-Arabidopsis          DSHQYNQLYKLVSTMGRLLNDIQGFKRESAEGKLNAVSLHMKHERDNRSKEVIIESMKGL
KS-rice                 RDHEYNELFRLMSTCGRLLNDSQGFERESLEGKLNSVSLLVHHSGGSISIDEAKMKAQKS
comp112296_c0_seq1      VDPDYLSIMNSISTISRISNDIRGYERESRQGKLSCVTLFMK-NNEVKKDMDAVLHFTSL
CPS-Arabidopsis         NFFTHTHFVRLAEIINRICLPRQ----------------YLKARRNDEKEKTI-----KS
CPS-rice                NNTEGSWFIQLASS---ICDSLHA---------------KMLLSQDTKKNETTINQIDKE
                              : .  .    :    .                 :       .

KS-Arabidopsis          AERKREELHKLVLEEKGSV-VPRECKEAFLKMSKVLNLFYRKDDGFTS-NDLMSLVKSVI
KS-rice                 IDTSRRNLLRLVLGEQGAV--PRPCKQLFWKMCKIVHMFYSRTDGFSSPKEMVSAVNAVV
comp112296_c0_seq1      RDTEMRKLTEKIIGQT---RFPRMFISIHLNMARIINFFYSKGDGHTSLDAMYEHVNNTL
CPS-Arabidopsis         MEKEMGKMVELALSESDTF---RDVSITFLDVAK--AFYYF---ALCG-DHLQTHISKVL
CPS-rice                IELGMQELAQYLLPRVDDRRINNKTKQTFLSIVK--SCYYA---ANCSPHMLDQHISEVI
                         :    ::    :         .     . .: .    :*    .  .   :   :. .:

KS-Arabidopsis          YEPVSLQKESLT-------
KS-rice                 KEPLKLKVSDPYGSILSGN
comp112296_c0_seq1      FRPIT--------------
CPS-Arabidopsis         FQKV---------------
CPS-rice                FEQVI--------------


Multiple sequence alignment of GID1 by MUSCLE (3.8)
comp106432_c0_seq1      MLQPAPLPPGHSPDSKGVVPLSTWVLISNFKLSYNLLRRPDGTFNRHLAEFLDRKVMANS
comp108403_c0_seq1      ------------------------------------------------------------
At3g05120-GID1L1-Arabido---MAASDEVNLIESRTVVPLNTWVLISNFKVAYNILRRPDGTFNRHLAEYLDRKVTANA
Os05g0407500-GID1-rice  ---MAGSDEVNRNECKTVVPLHTWVLISNFKLSYNILRRADGTFERDLGEYLDRRVPANA


comp106432_c0_seq1      SPVDGVASMDVLIERTTGVWGRIFWQ--AEHNTDQS----SKPL---------------I
comp108403_c0_seq1      --------MDVMIDRAIGLWGRLFWA--CETLADPAVRLRRQPL---------------L
At3g05120-GID1L1-ArabidoNPVDGVFSFDVLIDRRINLLSRVYRPAYADQEQPPSILDLEKPV--------DGDIVPVI
Os05g0407500-GID1-rice  RPLEGVSSFDHIIDQSVGLEVRIYRAA-AEGDAEEGAAAVTRPILEFLTDAPAAEPFPVI
                               :* :*:.  .:  *::.   .:     .     .*:               :

comp106432_c0_seq1      IYFHGGSFAHSSANSAIYDAMCRRLTKMCSVVILSINFRRAPENRYPCAYDDGITSMRWA
comp108403_c0_seq1      TYFHGGSFVHSSANSSIYDAMCRRLARMCGVVVLSVNFRRAPEHRFPIAYEDCAACVRWA
At3g05120-GID1L1-ArabidoLFFHGGSFAHSSANSAIYDTLCRRLVGLCKCVVVSVNYRRAPENPYPCAYDDGWIALNW-
Os05g0407500-GID1-rice  IFFHGGSFVHSSASSTIYDSLCRRFVKLSKGVVVSVNYRRAPEHRYPCAYDDGWTALKW-
                          :******.****.*:***::***:. :.  *::*:*:*****: :* **:*   .:.*

comp106432_c0_seq1      QGIHGSACLRSLGCDPQGRCFLAGDSSGGNIAHNVAVRAAEEGLPLSGFILLMPMFGGQA
comp108403_c0_seq1      KGAVGRQCLAEVGGDPD-RCFVAGDSSGGNIAHAVAVILAAEGVRLSGMVLLMPMFGGQQ
At3g05120-GID1L1-Arabido---VNSRSWLKSKKDSKVHIFLAGDSSGGNIAHNVALRAGESGIDVLGNILLNPMFGGNE
Os05g0407500-GID1-rice  ---VMSQPFMRSGGDAQARVFLSGDSSGGNIAHHVAVRAADEGVKVCGNILLNAMFGGTE
                                     *.. . *::********** **:  . .*: : * :** .****

comp106432_c0_seq1      RMPSEMALDGKYFVTLKDRDWYWRAFLPLGTSREHPACNPFSIHAPQLQRINLPPCLVVV
comp108403_c0_seq1      RTPAERLLDGKYFVSIKDRDWYWRAFLPPGATRDHPACDPFSPIAPSLVHLPLPPCLAVV
At3g05120-GID1L1-ArabidoRTESEKSLDGKYFVTVRDRDWYWKAFLPEGEDREHPACNPFSPRGKSLEGVSFPKSLVVV
Os05g0407500-GID1-rice  RTESERRLDGKYFVTLQDRDWYWKAYLPEDADRDHPACNPFGPNGRRLGGLPFAKSLIIV
                        *  :*  *******::.******.*:** .  *:****:**.  .  *  : :. .* :*
```

```
comp106432_c0_seq1      GGYDLLQDWQLRYVYGLKQAGKPVRVMFLEQATIGFFLLPNSDLFYSLVEELRTFLDAPR
comp108403_c0_seq1      GGYDILQDWQLRYVHSLQRAGKSVQLLFLEQATMGFFLLPNSDLFYTLVDRLKEFFGNP-
At3g05120-GID1L1-ArabidoAGLDLIRDWQLAYAEGLKKAGQEVKLMHLEKATVGFYLLPNNNHFHNVMDEISAFVNAEC
Os05g0407500-GID1-rice  SGLDLTCDRQLAYADALREDGHHVKVVQCENATVGFYLLPNTVHYHEVMEEISDFLNANL
                        .* *:  *.** *. .*.  *: *.::  *:**:**:****.  :: ::: :  *..

comp106432_c0_seq1              --
comp108403_c0_seq1              --
At3g05120-GID1L1-Arabidopsis    --
Os05g0407500-GID1-rice          YY

Multiple sequence alignment of GAI by MUSCLE (3.8)
At1g14920-GAI-ArabidopMKRDHHHH-----------HHQDKKTMMMNEEDDGNGMDELLAVLGYKVRSSEMADVAQK
AAX07462.1-GAI-rice   MKREYQEAGGSSGGGSSADMGSCKDKVMAGAAGEEEDVDELLAALGYKVRSSDMADVAQK
comp46913_c0_seq1     MLCCPSDS-----------TFSQRQSMGLGREAD---IEALLADAGYNVKASDLALVAQR
comp74927_c0_seq1     MFQSPSDS-----------LLPQNQTMGLG-DAD---IETLLAGAGYNVKASDLALVAQR
                      *                     ...:  .   :   :: ***  **:*.:*::* ***.

At1g14920-GAI-ArabidopLEQLEVMMS--------NVQEDDLSQLATETVHYNPAELYTWLDSMLTD-----------
AAX07462.1-GAI-rice   LEQLEMAMGMAGVSAPGAADDGFVSHLATDTVHYNPSDLSSWVESMLSE-----------
comp46913_c0_seq1     LEQLDSLCA--------SQDTGALSYLSSEAVHYNPSDMAAWLECMIGELGPSSVPGDVG
comp74927_c0_seq1     LELLDSLCS--------SHDAGALSYLSSEAVHYNPSDMASWLECMIGELAPSSAPTDIC
                      ** *:   .          : . :* *::::*****::: :*::.*: :

At1g14920-GAI-Arabidop------------------------------------------------------------
AAX07462.1-GAI-rice   ------------------------------------------------------------
comp46913_c0_seq1     GTQRPASENPLPPLSSTFYDFGNVNSSVPCSSVVKNSFIDQKSSVHSPFVDCPPKQAVPQ
comp74927_c0_seq1     SFQG-VLEGHFSQQTSGHYGIDDVYGPFGCTRGTDYQLNKPNTFLQDSFPNPQPKQGALP


At1g14920-GAI-Arabidop-----LNPP-------------------------------------------------
AAX07462.1-GAI-rice   -----LNAPLPPIPPAPPAARHASTSSTVTGG----------------------GGSGFFE
comp46913_c0_seq1     PALGILDPTAEGLPSISQLIKDAIGHNGGAPAAS---ATLKGYPGIALKDRTPGGLQQHK
comp74927_c0_seq1     SVL--LQTPVECVTSIPQLIRDAIGNQGGASATADRNESRSSYPGVTLPKRDVGGLHHYK
                           *:..

At1g14920-GAI-Arabidop------SSNAEYDL-------------------KAIPGDAILN-----------------
AAX07462.1-GAI-rice   LPAAADSSSSTYAL-----------------RPISLPVVATAD-----------------
comp46913_c0_seq1     IIEDQGSSNQVGAF----------FPRSSAGDPPQLSNMSTLQQAVPIPSPKMHGNPSLS
comp74927_c0_seq1     ELEDQGSCNQAKGFCAGNSTQPCLISHVSLQKSCSMPSLHQLQQAGHISATQARGSFSFH
                            *..    :                     :.     :

At1g14920-GAI-Arabidop-----------------QFAIDSASSSNQ--GGGGDTYTTNKRLKCSNG-----------
AAX07462.1-GAI-rice   -----------------PSAADSARDTKRMRTGGGSTSSSSSSSSSLGGGASRGSVVEAA
comp46913_c0_seq1     MQHQMQSQSLFSSVSIPPPNPASSQSSSNKVPRTGSPSPVHVQRQCHRPPQNQGTVRTST
comp74927_c0_seq1     TQHQTQGQSFSSPAA--SPATTSSQNSNN--KATYHEAPSVRFQQQLHRKVNQEEVKITE
                                       *: .:..         .     .

At1g14920-GAI-Arabidop--VVETTTATAESTRHVVLVD---------------------------------------
AAX07462.1-GAI-rice   PPAMQGAAAANAPAVPVVVVD---------------------------------------
comp46913_c0_seq1     AMVMASVSPSNSSPVSISYQDHSSPHDKEASYVHIQSPSAKRTRSQTVHECPYDDISNDE
comp74927_c0_seq1     PEVTADLSPSSSSPMSVSYQEHCSPQDKDSIY-HMRYAPSKHANSQTMQTCPYTEVVDYE
                        .    :.:  ..  :   :

At1g14920-GAI-ArabidopSQENGVRLVHALLACAEAVQKENLTVAEALVKQIGFLAVSQIGAMRKVATYFAEALARR
AAX07462.1-GAI-rice   TQEAGIRLVHALLACAEAVQQENFAAAEALVKQIPTLAASQGGAMRKVAAYFGEALARR
comp46913_c0_seq1     NAQESGIKLVHLLMACAEAIQNDELAAAVDMVREIKRLASCTSGAMSKIASYFAESLSQR
comp74927_c0_seq1     NVQESGIKLVHLLMACAEAIQNNALAAAVDMVREIKRLASSTRGTMSKVANYFVESLARC
                        ** *:.*** *:*****:*:: ::.*  :*.:*  ** .  *:* *:* ** *:*:.

At1g14920-GAI-ArabidopIYRLSPSQSPI---DHSLSDTLQMHFYETCPYLKFAHFTANQAILEAFQGKKRVHVIDFS
AAX07462.1-GAI-rice   VYRFRPADSTLL--DAAFADLLHAHFYESCPYLKFAHFTANQAILEAFAGCRRVHVVDFG
comp46913_c0_seq1     IYPASKDNWARIYEAEAVSEMLYASFYEACPYLKFAHFTANQAILEAFQGHKVVHIIDFN
comp74927_c0_seq1     IYPGNKCDWAYLCQADALSELLYANFYEALPYLKFAHFTANQAILEAFQGHKFVHIIDFN
                      :*     : .      :.:: *   ***: ***************** * . **::**.

At1g14920-GAI-ArabidopMSQGLQWPALMQALALRPGGPPVFRLTGIGPPAPDNFDYLHEVGCKLAHLAEAIHVEFEY
AAX07462.1-GAI-rice   IKQGMQWPALLQALALRPGGPPSFRLTGVGPPQPDETDALQQVGWKLAQFAHTIRVDFQY
comp46913_c0_seq1     LMQGSQWPELIKALAVRSEGPPHLRMTGIGPPRPDNKDVLQEVGVKLAELAGSVNVEFSF
comp74927_c0_seq1     LMQGSQWPALIQALADREEGPPYLRMTGIGLPHQDNKDVLQEVGKELAELAHSVNVKFSF
                      : ** *** *::*** *  *** :*:**:* *  *: * *::** :** :* ::.*.*.:
```

```
At1g14920-GAI-ArabidopRGFVANTLADLDASMLELR-----PSEIESVAVNSVFELHKLL-------GRPGAIDKVL
AAX07462.1-GAI-rice   RGLVAATLADLEPFMLQPEGEADANEEPEVIAVNSVFELHRLL-------AQPGALEKVL
comp46913_c0_seq1     RGMVAAKLDDVKPWYFEVK-----PG--EAIAVNSILQMHRLLYGHVASDPSKALIDEVL
comp74927_c0_seq1     RGMVATKLEDVKPWYFEVN-----PG--EAIAVNSILQMHRLLYGCVGSDPSKAPIDEVL
                      **:** .* *:..  ::           * :****::::*.**          . :::**

At1g14920-GAI-ArabidopGVVNQIKPEIFTVVEQESNHNSPIFLDRFTESLHYYSTLFDSLEGVPS------------
AAX07462.1-GAI-rice   GTVHAVRPRIVTVVEQEANHNSGSFLDRFTESLHYYSTMFDSLEGGSSGQAELSPPAAGG
comp46913_c0_seq1     SSIKSLNPKVVTVVEQEANHNSNMFLERFVEALHYYSTMFDSLEASSLDPL---------
comp74927_c0_seq1     SFIKSLKPKVVTLVEQEANHNGSIFLERFVEALHYYSTMFDSLEASSLDPQ---------
                      . :: :.* :.*:****:***.  **:**.*:******:*****. .

At1g14920-GAI-Arabidop--GQDKVMSEVYLGKQICNVVACDGPDRVERHETLSQWRNRFGSAGFAAAHIGSNAFKQA
AAX07462.1-GAI-rice   GGGTDQVMSEVYLGRQICNVVACEGAERTERHETLGQWRNRLGRAGFEPVHLGSNAYKQA
comp46913_c0_seq1     --GPEMVCSEMYLGREIANIVAREGAERVERHEPLSAWRKRMSNAGFKQVHLGSNAFDQV
comp74927_c0_seq1     --SSEMACAEAYLAREITNVLACEGAERVERHEPLSQWRKRMSNAGFKPLHLGSNAFNKV
                        . : . :* **..:* *::* :*.:*.****.*. **:*:. ***   *:****:.:.

At1g14920-GAI-Arabidopsis      SMLLALFNGGEGYRVEESDGCLMLGWHTRPLIATSAWKLSTN
AAX07462.1-GAI-rice            STLLALFAGGDGYRVEEKEGCLTLGWHTRPLIATSAWRVAAA
comp46913_c0_seq1              SYMLKYFS-GEGYTVEENRGCLTLGWHNRPLIAASAWECG--
comp74927_c0_seq1              SVLLKVFS-GEGYTVEENKGCLTLGWHNRPLIASSAWQCG--
                               * :*  *  *:** ***. *** ****.*****:***  .
```

Multiple sequence alignment of MYBs by MUSCLE (3.8)

```
comp37605_c0_seq1       --MGTVERSDSRHRGERMTAC-----------------------------------EMRR
comp82703_c0_seq1       ----MESRSTRRHFQARQPPSLQQRENVP---------------------------SLKK
comp91285_c1_seq1       MAAEAAKSSGNGKAGGAEGADVDEDCSDSSC----GKRDKHVECAAGTSVQTRGRKEMRK
comp106205_c0_seq1      MEDEGCQQSVARTSGAISGVVMAKGASNFSSEKGHGVGSSHSGSEGDGELCQVG--SLRK
AT5G06100.2-myb33       ---MSYTSTDSDHNESPAADDNGSDC---------------RSRWDGH-------ALKK
PpGAMYB1_protein        ----MDLSSDIGQDGG-----------------------------------------SLKK
PpGAMYB2_protein        ----MDMSSDVGLDGG-----------------------------------------ALKK
                                :                                                :..

comp37605_c0_seq1       GTWTPEEDELLMAYVEKHGASAWNMAPFYYPELRRTGKSCRLRYTNQLRPGIRRHPVSPE
comp82703_c0_seq1       GPWTAEEDALLLAYVNQHGNGNWNSVQ-KFSGILRCGKSCRLRWTNHLRPHLKKCSFSRE
comp91285_c1_seq1       GPWSAWEDQLLLDYVSKHGKGNWKEVA-QRSGLRRCGKSCRLRWTNQLRPNLRKDRFTPA
comp106205_c0_seq1      GPWSPWEDELLLQYVRKHGQGNWKEVA-RRSGLRRCGKSCRLRWTNQLRPNLRKDRFTPS
AT5G06100.2-myb33       GPWSSAEDDILIDYVNKHGEGNWNAVQ-KHTSLFRCGKSCRLRWANHLRPNLKKGAFSQE
PpGAMYB1_protein        GPWTSAEDSILISYVTKHGEGNWNSVQ-KHSGLFRCGKSCRLRWANHLRPNLKKGAFTPE
PpGAMYB2_protein        GPWTSAEDSILISYVTKHGEGNWNSVQ-KHSGLFRCGKSCRLRWANHLRPNLKKGAFTPE
                        *.*:. ** :*: ** :** . *: .    . : * *******::*:*** :..  .:

comp37605_c0_seq1       ELLLILRLHSQYGNQWSKIASMVPGRTDNSVKNIVNMHLKKARRRAATLGLARAAAAAAA
comp82703_c0_seq1       EERLIIDQHAAIGNRWSRIAAMLPGRTDNEVKNFWNTRVKRLLRAGKPLYPPDIIPMVQA
comp91285_c1_seq1       EEATILLLHSIHGNKWAKISAQVPGRTDNSIKNFMNTRAKRQRR------------QTAA
comp106205_c0_seq1      EEATILLLHSIYGNKWAKISAQVPGRTDNSIKNFMNMRAKRQKR------------QSAA
AT5G06100.2-myb33       EEQLIVELHAKMGNRWARMAAHLPGRTDNEIKNYWNTRIKRRQRAGLPLYPPEMHVEALE
PpGAMYB1_protein        EERTIVELHAKLGNKWARMAAQLPGRTDNEIKNYWNTRIKRRMRAGLPVYPAE---KAKS
PpGAMYB2_protein        EERIIVELHAKIGNKWARMAAQLPGRTDNEIKNYWNTRIKRRMRAGLPVYPAD---KSKS
                        *   *:  *:  **.*:.::: :******.:**  * . *.  *

comp37605_c0_seq1       QLLLPSSTQP----------VCVGPSSTSVSDAVETPSQLFPF------SSHGLQALAGV
comp82703_c0_seq1       RLGQREPQQFVNERPANLGVVGDGRDDKSIGGKTSKFISTITQDQGTSSPGGNTRHVQIN
comp91285_c1_seq1       -----APSLH--------------QRISAAEDPACH-------------PGTSKFSLLHD
comp106205_c0_seq1      -----SSVQP-------------RRTSTSL-------------------PISSALELDAA
AT5G06100.2-myb33       -----WSQEY-------------AKSRVMGEDRRHQ---DFLQ------LGSCESNVFFD
PpGAMYB1_protein        -----SPTQY-------YGKPSDGRSFISGEDADCDFISSFPQ------PGDHLHNLHAI
PpGAMYB2_protein        -----SPTQY-------YVEPSDGRSFISGKDADCDFTSSFAQ------PGDHLHNLHVS
                              .                                                  :


comp37605_c0_seq1       NS----------------------LHTCGGWD------ATSSQLPSVRKCQSSSLSPEL-
comp82703_c0_seq1       PTVSKALLTLGGQINGSSKSVQDSLFYLGNGDNTVDLGVYENSSPALRCSNEITLDSNMI
comp91285_c1_seq1       LE-----------------------GVSSSSS------VMSFRLNAT-------------
comp106205_c0_seq1      SA----------------------SSSSSSNS------RPSSRVHPT-------------
AT5G06100.2-myb33       TLNFTDMVPGTFDLADMTAYKN--MGNCASSP------RYENFMTPTIPSSKRLWESEL-
PpGAMYB1_protein        NECDHSLTRNCLSMIPMGALAPS-TTNQGTNQ------VLNKTIGNPFEQIEVFRNSHHG
PpGAMYB2_protein        NPRDSRLMRNCFSITSMSELAPSATTNQGTNH------VLNKSMGNPFEQTEVFRNSQHG
                                              .          .
```

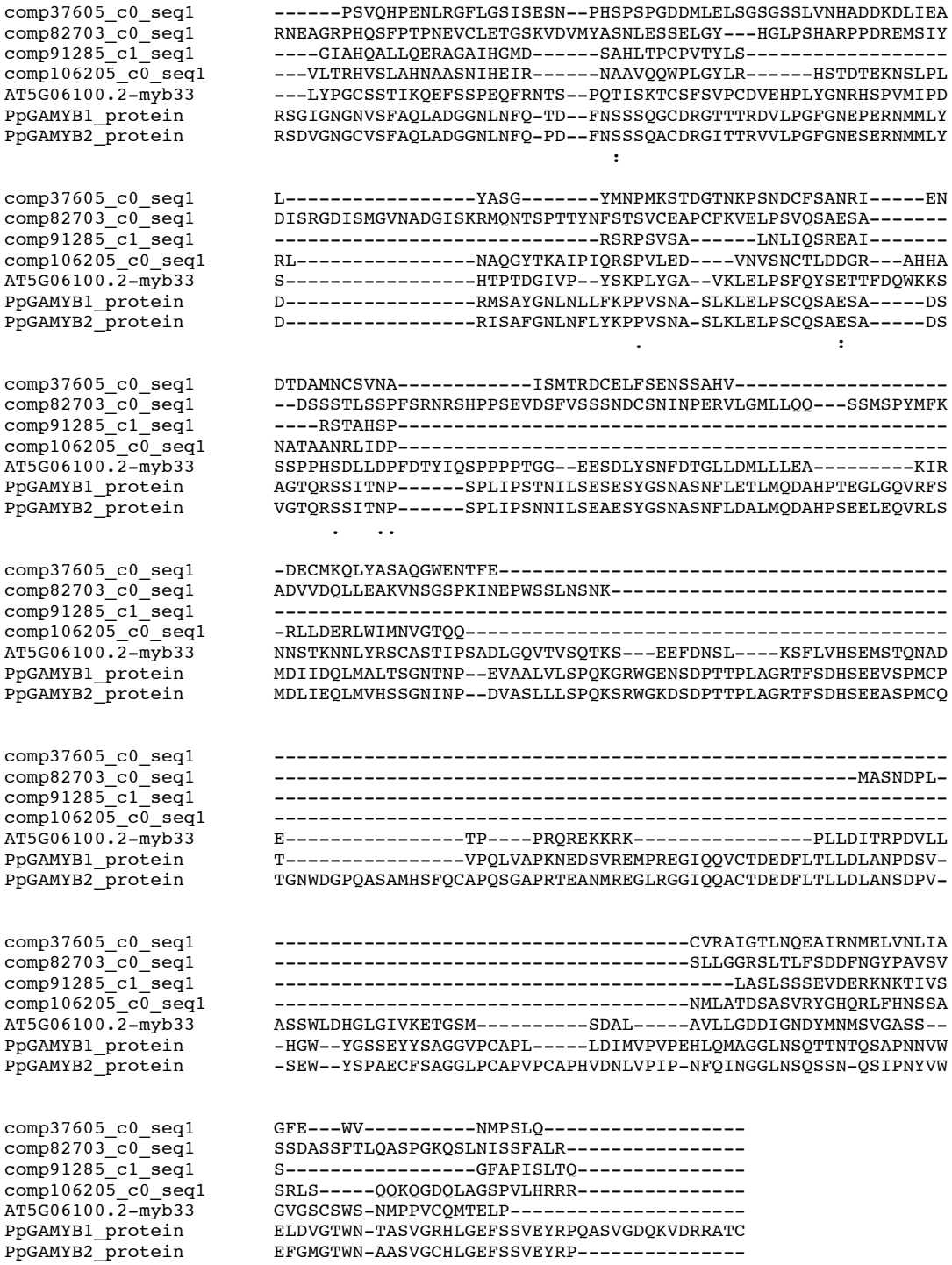


Figure 2.7. Alignments of *CPS/KS*, *GID1* and *GAI* genes from *Ceratopteris*, rice and *Arabidopsis*, plus an alignment of *MYB* genes from *Ceratopteris*, *Physcomitrella* and *Arabidopsis*.

CHAPTER 3. CHARACTERIZATION OF TRANSCRIPTIONAL COMPLEXITY DURING EARLY GAMETOPHYTE DEVELOPMENT USING RNA-SEQ

3.1 Introduction

The prior RNA-Seq experiment provides insight into the molecular and genetic mechanisms controlling sex determination in Ceratopteris. The experiment in the previous chapter lays the foundation for a larger time-course experiment, which will provide an even more complete transcriptome assembly and the ability to observe the transcriptional landscape early in gametophyte development. Time-points have been chosen based on the stages in Ceratopteris gametophyte development.

Six distinct stages of gametophyte development have been characterized in Ceratopteris and are described in (J. A. Banks, L. Hickok, & M. A. Webb, 1993a). Stage 1 begins when the spore is inoculated into media. The spore, with spore wall still intact, begins to imbibe water. Sequencing was performed on 0 DAI (dry spore) samples, prior to stage 1 for a couple of reasons: first, sequencing the dry spore samples provides insight into which transcripts are stored in the spore prior to germination, and second, to provide a baseline with which to compare other time-point data. For example, prior to performing sequencing on dry spores, we have not had the information to be able to conclude whether differentially expressed genes that are more highly expressed in $+A_{CE}$ samples at 4.5 DAI are up-regulated in the presence of A_{CE} or down regulated in the

absence of A_{CE}. Stage 2 is at 3-4 DAI, when the spore wall cracks; it is in stage 2 that the gametophyte becomes competent to respond to the male-inducing effects of A_{CE}. Exposure of the gametophyte to A_{CE} during stage 2 and onward is imperative for male gametophyte development. It is during stage 2, at 3 DAI that the tissue for the second time-point was harvested; after harvesting the tissue, A_{CE} was added to half of the remaining samples so that gene expression could be compared with and without A_{CE} (Banks et al. 1993b). Sequencing performed on samples collected at 3.5 DAI - 12 hours after A_{CE} was added, will hopefully lead to detection of genes that are early responders to A_{CE}. At 4-5 DAI, stage 3 of gametophyte development begins; at this stage the gametophyte consists of 3-5 cells and 1-3 rhizoids. Gametophytes lose competence to respond to A_{CE} at around 5 DAI. Samples collected at 4.5 DAI were sequenced in hopes of detecting gene expression differences that occur just before gametophytes lose competence to respond to A_{CE}. Two-dimensional growth begins in stage 4, 5-6 DAI. Thus, in hopes of detecting expression differences that occur just before male and hermaphrodite gametophytes become morphologically distinct, sequencing was performed on samples collected from gametophytes 5.5 days after inoculation. Male and hermaphroditic gametophytes become morphologically distinct in stage 5, 6-7 DAI and are sexually mature by stage 6, 10-12 DAI (Banks et al., 1993b).

These developmental stages and changes are the result of carefully orchestrated transcription of genes involved in developmental processes. However little is known about what these genes are and how dynamic the transcriptome is in early gametophyte development. Although studies of global gene expression in development across time have been performed in plants (Xu, Gao, & Wang, 2012; Zenoni et al., 2010), in general

little is known about the complexity of gene expression patterns in early development. Furthermore, many of the studies conducted have been on whole organs and are thus limited by the compound nature and complexity of plant organs and tissues(reviewed in (Schnable, Hochholdinger, & Nakazono, 2004)); RNA-Seq on fern gametophytes provides an opportunity to observe gene expression in the comparatively simple, haploid fern gametophyte in a time-course design, during early development. Moreover, very little is known about gene expression across development in the gametophyte. The fact that fern gametophytes are independent of the sporophyte gives a truly unique opportunity to study gametophyte development; studies on gametophytes in angiosperms are much more difficult due to the reduced nature of the angiosperm gametophyte (reviewed in (Banks, 1997a)). Although a few studies using massively parallel sequencing to observe transcriptomics of gametophytes have been performed (Aya et al., 2015; Chettoor et al., 2014; Loraine, McCormick, Estrada, Patel, & Qin, 2013; S. S. Wang et al., 2014), none have been performed which allow observation of gene expression across several time-points. Previously, a small gene expression analysis study using a microarray representing just over 3,000 genes was conducted over the first two days of Ceratopteris development. This study found that vast changes in gene expression take place within the first 48 hours after spore inoculation, and also found significant overlap between genes expressed during spore germination and genes expressed during angiosperm seed germination (Salmi, Bushart, Stout, & Roux, 2005). While highly informative, this microarray study on a small fraction of the total number of genes in the Ceratopteris genome does not address how the global transcriptome changes across time.

The experiment described in this chapter looks at global gene expression using RNA-Seq and is, to our knowledge, the first global gene expression time-course on gametophytes. The present study identifies genes and gene ontology (GO) terms likely to be important in germination and in early gametophyte development and for the first time examines just how dynamic the transcriptome of the young gametophyte is.

3.2 Materials and methods

3.2.1 Plants and growth conditions

Hn-n is the wild-type strain of *Ceratopteris richardii* used in this study, the origins of which are described in (L. G. Hickok, T. R. Warne, & M. K. Slocum, 1987). Gametophytes are cultured in -A_{CE} media referred to as fern media, or FM and described in (Banks, 1993) or cultured in +A_{CE} media referred to as conditioned fern media, or CFM media as described in (Banks, 1993). Spores were surface sterilized as described in (Banks, 1994a).

A repeated-measures design was used and time-points were carefully chosen based on developmental milestones(Banks et al., 1993b). Samples were grown in CFM media and harvested at 0, 3, 3.5, 4.5, and 5.5 days after inoculation. A_{CE} was added to half of the samples harvested at 3.5, 4.5, and 5.5 days after inoculation; gametophytes were either treated with A_{CE}, or not treated with A_{CE} beginning at 3 days. Three biological replicates were sequenced at each time-point, for each condition, however one dry spore sample generated very few useable reads, and thus this sample is excluded from all downstream analyses (Table 3.1).

3.2.2 Library preparation and sequencing

Harvested tissue was frozen and ground under N_2(l) until no intact cells were observed upon looking at tissue under a light microscope (for 30-60 minutes). Total RNA was then extracted using the RNeasy Plant Mini Kit (Qiagen, CA) and treated with DNase using the DNA-Free RNA Kit (Zymo Research, CA). Libraries were generated for all samples using the TruSeq Stranded mRNA Sample Prep Kit (Illumina, CA), were amplified using ten cycles, and fragmented for four minutes. Libraries were qPCR quantified, pooled in equimolar concentration, and paired-end strand-specific sequencing was performed on an Illumina HiSeq2000 platform at the Purdue Genomics facility.

3.2.3 Quality control and transcriptome assembly

To ensure that only high-quality reads were utilized in the analyses, quality control was performed using a number of available programs. The program clean_adapter.pl version 1.4 (Gribskov, pers. comm.) was used to remove Illumina adapter sequences. Trimmomatic version 0.30 (Lohse et al., 2012b) was utilized to trim reads based on quality score; bases with a quality score less than 20 were removed and reads that were under 30 bases in length post-trimming were removed. In order to remove reads mapping to contaminants, DeconSeq version 0.4.3 (Schmieder & Edwards, 2011a; Schmieder, Lim, & Edwards, 2012a) was run on each of the FASTQ files to remove reads aligning to chloroplast RNA, mitochondrial RNA, rRNA, viral, and bacterial databases; an identity threshold of 75 and a coverage value of 50 were used. The *de novo* assembly program Trinity, which uses a fixed k-mer size of 25 (release 2013-08-14) (Grabherr et al., 2011a), was used to assemble a transcriptome from the

FASTQ files. The program getpairs.pl (Gribskov, pers. comm.) was used to separate reads in FASTQ files into paired and unpaired reads.

3.2.4 Time-wise differential expression analysis

The program RSEM version 1.2.0 (B. Li & Dewey, 2011; B. Li et al., 2010) was used to align reads to the assembled transcriptome and to estimate expression levels of genes. DESeq2 (Love, Huber, & Anders, 2014) was used to identify differentially expressed genes across time using a Benjamini-Hochberg (Hochberg & Benjamini, 1990) corrected FDR of 5%. In order to reduce the number of hypothesis tests performed by DESeq2, a reference transcriptome was used in the differential expression analysis. To prepare the reference transcriptome, first a tBLASTn search (E-value cutoff of 10^{-10}) was performed of the new assembly against the *Pteris vittata* transcriptome (unpublished data), a BLASTn search against GenBank Ceratopteris ESTs (E-value cutoff of 10^{-20}), and a BLASTn search against the Ceratopteris transcriptome from Chapter 2 (E-value cutoff of 10^{-20}). Finally a BLASTx was run to compare the new assembly versus version 9.1 of the Phytozome protein database (Goodstein et al., 2012) using an E-value cutoff of 10^{-10}. Transcript assemblies without BLAST matches, as well as sequences with counts less than 0.3CPM (counts per million) were removed.

The design formula specified in DESeq2 allowed us to look for genes expressed differentially as a function of time. A differential expression analysis using the Wald test (Wald, 1943) was performed for each pair of consecutive time-points. In the Wald test, a beta prior is applied to moderate effect sizes from the GLM. These effect sizes are then used to calculate the p-value that the effect is different from zero (Wald, 1943). The β_{ir}

coefficient estimate for each gene is divided by its standard error and is compared to a normal distribution to determine whether the null should be rejected (Love et al, 2014). An additional biological significance fold-change cutoff of 2 was applied in selecting differentially expressed genes. A likelihood ratio test (LRT) (Neyman and Pearson, 1928) was also performed to test for differential expression across all time-points. An LRT is used to test multiple terms at once using a full and reduced model and is conceptually similar to ANOVA. In the LRT, both a full and a reduced model in which time has been removed were specified (Neyman and Pearson, 1928). The LRT is based the likelihood ratio, or, in the case of DESeq2, the log-likelihood ratio (Love et al, 2014), comparing both the reduced and full models and allowing calculation of a p-value (Neyman and Pearson, 1928, Love et al, 2014),. Thus, using the LRT we tested the null hypothesis that there is no effect of time.

3.2.5 Expression analysis validation with qRT-PCR

Tissue was grown and RNA extracted as described above for the RNA-Seq library preparation. Total RNA was treated with DNase using the DNA-Free RNA Kit (Zymo Research, CA), and was reverse transcribed into single-stranded cDNA using the Tetro cDNA Synthesis Kit (Bioline, MA). Approximately 1.5 ng cDNA was used as template for each qRT-PCR reaction, performed using the StepOne Real-Time PCR System (Applied Biosystems, NY) and the SYBR green PCR Master Mix from Applied Biosystems. PCR conditions were: 1 cycle of 20 minutes at 95°C, 40 cycles of 3 seconds at 95°C and 30 seconds at 60°C. Melt curves (15 seconds at 95°C, 60 seconds at 60°C, and 15 seconds at 95°C) were performed. All oligonucleotide primers were

900nM, and only those producing a single Tm peak were used. Three biological replicates of both $+A_{CE}$ and $-A_{CE}$ samples were performed for each template and three technical replicates were performed for each sample. Measurements were normalized to the amount of *CrEF1α* (GenBank accession number BE642078) transcript in the samples. The ΔC_t method was used in calculating relative fold changes (Livak & Schmittgen, 2001). The primer sequences used are listed in Table 2.1 in Chapter 2.

3.2.6 Annotation and assembly validation

To validate the assembly, first a BLASTn search was utilized to compare all predicted transcripts with read support in the assembly described in Chapter 2 with a database made from the new assembly; an E-value cut-off of $1x10^{-20}$ was used. Next, tBLASTn was used to compare all Arabidopsis proteins to the database of the Ceratopteris transcripts (E-value < 10^{-10}). Then, tBLASTx (E-value < 10^{-10}) was used to compare for similarity to Lygodium predicted proteins from the assembled transcriptome of *Lygodium japonicum* (Aya et al., 2014) to a database of the Ceratopteris predicted transcripts. A BLASTx search (E-value < 10^{-10}) of Arabidopsis ultra-conserved orthologs (Kozik et al., 2008) against the Ceratoperis assembly was used to estimate the number of genes sequenced, as has been done in other studies(Der et al., 2011; L. Jiang et al., 2013; Kozik et al., 2008; Y. Wang et al., 2012). Additionally we compared the current Ceratopteris assembly with 5,133 publicly available Ceratopteris ESTs downloaded from GenBank using a BLASTn search with an E-value cut-off of 10^{-20}. Additionally, MEGAN5 (Huson et al., 2011) was used to perform a taxonomic analysis on the transcripts used in the differential expression analysis. The program MEGAN

(MEta Genome ANalyzer) allows for functional and taxonomic analysis and characterization of sequence datasets (Huson et al., 2011). The XML files used as input into MEGAN5 were obtained from a BLASTx search using sequences from the reference transcriptome (the predicted sequences with similarity to known sequences as well as at least 0.3CPM reads aligning) as queries and searching against the nr database (E-values<10^{-10}). The default parameters were used in performing the analysis. Thus any sequence hits that have a bit score less than 90% of the value of the best hit's bit score were ignored. Due to the LCA-algorithm used by MEGAN5 some nodes have no sequences assigned to them. This is due to the fact that in MEGAN5, if sequences match two nodes A and B, and A is an ancestor of B, the sequence is assigned only to node B. Node labels with zero sequences assigned were deleted to enhance readability.

A variety of methods were used in annotating the transcriptome. A BLASTx search was performed using the transcriptome assembly as the query and the TAIR10 protein database as the subject (using an E-value cutoff of 10^{-10}). RepeatMasker (Chen, 2004) was used to identify repetitive sequences in the transcriptome. Protein-encoding, differentially expressed genes were annotated using the Trinotate workflow (Ashburner et al., 2000; Finn et al., 2011; Grabherr et al., 2011b; Kanehisa et al., 2012) using the version released on 2014-02-25, with a 100 amino acid minimum cutoff for ORFs. BLAST2GO (Aparicio et al., 2006; Conesa & Gotz, 2008; Conesa et al., 2005; Gotz et al., 2008) was run to map GO terms to sequences and to make multi-level pie charts. InterproScan was utilized to perform a protein functional analysis (P. Jones et al., 2014; Quevillon et al., 2005; Zdobnov & Apweiler, 2001).

3.2.7 Unsupervised clustering

Unsupervised clustering was performed to group genes based on expression profiles. After the differential expression analysis between pairs of consecutive time-points was performed, the $\log_2$(fold-changes) and adjusted p-values were used to assess the expression pattern of each predicted transcript, across all time-points. A negative $\log_2$(fold-change) indicates that a predicted transcript is decreasing in expression between time t and t+1, whereas a positive $\log_2$(fold-change) indicates that a predicted transcript is increasing in expression between time t and t+1. To be considered significantly different between times t and t+1, the adjusted p-value had to be less than 0.05. In R, a matrix containing a row for each predicted transcript and a column for each pair of consecutive time points (0-3DAI, 3-3.5DAI, 3.5-4.5DAI, and 4.5-5.5DAI) was made. Each element in the matrix was filled with a "0" representing no significant change in expression, "1" representing an increase in expression between time t and t+1, or "-1" representing a decrease in expression between time t and t+1. Predicted transcripts having the same expression trends were grouped together in clusters. Thus, all transcripts that did not change significantly in expression across time had associated entries in the matrix of "0", "0", "0", "0", and all of these transcripts were grouped together. Overall, 4 comparisons were made and so there were 3^4=81 clusters that transcripts could be grouped into. Clusters containing 500 transcripts or more were analyzed further to identify enriched functional categories within the clusters using GOSeq (see below for details) (Young et al., 2010). Additionally clusters with expression patterns that were deemed to be biologically interesting were analyzed further for functional enrichment.

3.2.8 Enrichment analysis

Enrichment analyses were performed using GOSeq v. 1.18.0 (Young et al., 2010) to identify overrepresented GO terms amongst the differentially expressed genes, genes expressed at each time-point, and amongst various clusters of genes. GOSeq is designed specifically for performing GO enrichment analyses on RNA-Seq data and takes into account length bias when performing the analyses. GOSeq uses a probability weighing function (PWF) to quantify how the probability of a differentially expressed gene changes with respect to its length. The PWF is calculated by fitting a cubic spline with a montonicity constraint to the differential gene analysis data, with a "0" representing a gene that is not differentially expressed and a "1" representing a gene that is differentially expressed. The PWF then forms the null hypotheses for the enrichment test. GOSeq then calculates P-values for each GO category using a resampling technique (Young et al., 2010). In addition to GOSeq taking length bias into account, the package does not impose cutoffs based solely on the number of times a term appears and thus even GO terms that are very specific and thus less abundant compared to other more general (and less useful GO terms) can be included in the results, provided they are statistically significantly enriched (Young et al., 2010). The GO terms mapping to the whole Ceratopteris transcriptome were used as a reference and a 5% FDR was used (Hochberg & Benjamini, 1990). A 5% FDR was chosen in order to keep the risk of making a Type I error relatively small, while attempting to not making the area in which one rejects the null so small that we miss to identify many truly differentially expressed genes. Enrichment analyses were performed to test for enriched GO terms and also enriched Plant GOslim terms.

3.3 Results and discussion

3.3.1 RNA-Seq and *de novo* assembly of the Ceratopteris transcriptome

Overall a total of ~3.4 billion reads, each ~100bp in length were sequenced. With the exception of one dry spore sample which generated very few useable reads, and thus was excluded from all downstream analyses, each developmental stage and condition had 3 biological replicates sequenced (Table. 3.1). Each developmental stage was represented by at least 336 million reads (Table 3.2). Several programs were run to filter and trim reads (Table 3.3). Reads were overall of high quality, and only 6% of reads were removed during filtering and trimming. The program Trinity (Grabherr et al., 2011a) was used to assemble the RNA-Seq reads into a transcriptome assembly, containing 395,694 sequences and 309,910 subcomponents with an N50 of 1,170 bases and an average sequence length of 713 bases (Table 3.4). Overall, ~89% of reads were aligned to the 395,694 predicted transcripts and counted using RSEM (B. Li & Dewey, 2011; B. Li et al., 2010). A total of 339,372 sequences had read support, though many of these sequences had very few reads align.

The removal of sequences with no read support or with very low counts is now a common practice in RNA-Seq differential expression analyses (Rau, Gallopin, Celeux, & Jaffrezic, 2013). Filtering in this manner is particularly useful when a *de novo* assembly has been performed, as many sequences are generated, some of which are no doubt lowly expressed transcripts without annotations and were thus not going to be followed up on experimentally. Removing such sequences can greatly improve the power to detect differentially expressed genes. However we wanted to choose a filtering criterion wisely, particularly so that lowly expressed genes are not filtered out; we do not want to lose

biologically meaningful data. One suitable filtering criterion is to filter based on counts per million (CPM) (M. D. Robinson et al., 2010). In order to determine an appropriate cutoff to use, a graph of the coefficient of variation versus the average counts normalized for library size of the genes (the baseMean) (Love et al., 2014) (Figure 3.1.), was generated. In this graph we observe a well-known phenomena – genes with very low counts have variable and often large coefficients of variation. Thus, we would likely not be able to detect truly differentially expressed genes at this level. Furthermore, at very low levels of expression, downstream wet lab operations are challenging – qRT-PCR simply cannot detect such low levels of expression and cloning such lowly expressed transcripts is difficult. The coefficient of variations begins to smooth out around a baseMean of 15-20 (Love et al., 2014). Based on the calculated library sizes, a baseMean of 15-20 counts corresponds to ~0.3CPM and thus 0.3CPM was selected as the filtering cutoff. Thus, in order to not be filtered out of the analysis, a gene had to have an average across samples of at least 0.3CPM in at least one time-point.

3.3.2 A reference transcriptome was prepared using read count data and sequence similarity

As mentioned previously, 339,372 sequences from the Trinity assembly had reads align in RSEM. A reciprocal BLASTn was performed using the transcriptome assembly described in Chapter 2 and the time-course transcriptome assembly. Overall, 139,227 sequences out of 147,117 (94.6%) sequences from the assembly described in Chapter 2 had BLAST hits with E-value $<10^{-20}$ to sequences in the new time-course transcriptome assembly. A total of 153,561 sequences out of 373,717 sequences (~41%) in the time-

course assembly had hits with E-value < 10^{-20} to sequences in the transcriptome assembly previously described in Chapter 2. Thus, the newer assembly has many more assemblies. Although Trinity is quite successful in reconstructing transcriptomes from short reads, it is well-known that as the number of reads included in the transcriptome assembly increases, the number of contigs assembled also increases (J. Zhang, Ruhlman, Mower, & Jansen, 2013). The large number of transcripts could potentially lead to a loss of power in the differential expression analysis. Therefore, in order to address this concern, both a 0.3CPM cutoff and annotation were utilized to create a "reference assembly", thereby reducing the number of sequences in the assembly to a more manageable set of transcripts.

To prepare a reference transcriptome, first sequences were removed which failed to meet the 0.3 CPM cutoff, leaving 66,925 sequences (including isoforms) (Table3.5). After filtering out sequences without BLAST similarity as well as sequences with counts <0.3CPM, with 42,798 sequences (including isoforms) and 32,128 subcomponents, which were considered to be genes, due to the results of a recent study published by Navidson and Oshlack in 2014 (Table 3.4). Navidson and Oshlack found that the clustering information provided by Trinity, in which subcomponents are utilized in downstream analyses as 'genes' and sequences as 'isoforms' is quite accurate (Navidson & Oshlack, 2014). A total of 86.36% of reads aligned to the reference. Thus, overall, running RSEM on only the transcripts included in the reference transcriptome made little difference in the number of reads aligning.

3.3.3 Transcriptome assembly and coverage assessment

In order to assess the completeness of the transcriptome, Arabidopsis "ultra-conserved orthologs" (Kozik et al., 2008), were used to estimate the number of genes sequenced, as has been done in other studies (Der et al., 2011; L. Jiang et al., 2013; Y. Wang et al., 2012). Out of 357 of these single-copy genes conserved amongst Eukaryotes, similar sequences to 100% of these sequences were detected using BLASTx (E-value < 10^{-10}). Additionally the Ceratopteris assembly was compared to 5,133 publicly available Ceratopteris ESTs using BLASTn, most of which were obtained from developing gametophytes (Salmi et al., 2005); overall, 4,976 (~97%) had hits with E-value <10^{-20} and 4475 (87%) had hits with an E-value of 0. Moreover, the Ceratopteris assembly was compared by tBLASTx to the transcripts from a recently assembled transcriptome of *Lygodium japonicum* (Aya et al., 2014). Out of 37,676 transcripts, some of which have been shown to be specific to the sporophyte generation, 25,555 sequences (67%) have similarity to genes in the Ceratopteris assembly. A tBLASTn comparison of the TAIR10 protein sequences to the Ceratopteris reference transcriptome shows that 26,947 out of 35,386 (76%) of the protein sequences have similarity to sequences in the Ceratopteris reference transcriptome assembly. Based on these measures, it seems that a large percentage of the transcriptome has been successfully sequenced and assembled.

A cladogram (Fig 3.2) was made using MEGAN5 to allow quantification of contaminants in the reference transcriptome (Huson et al., 2011). The nodes in the tree are proportional to the number of sequences assigned to them. MEGAN uses the LCA (lowest common ancestor) algorithm to assign reads only to a taxonomic level that can be inferred with confidence. The "assigned" number, shown next to the taxonomic names,

is the number of sequences that through a BLASTx search are assigned only to that node and not to any children/grandchildren of that node (Huson et al., 2011). Very few contaminants were observed in the cladogram (only 391 sequences) and the vast majority of the sequences (31,366) were assigned to Viridiplantae, thus the decontamination of reads was successful and the transcriptome does not show significant levels of contamination sequences.

3.3.4 Functional annotation of the Ceratopteris assembly

The reference transcriptome was annotated using several methods. Overall, out of 42,798 sequences, 33,880 (79%) had BLASTx hits to the non-redundant database and 29,284 (68%) had BLASTx hits to the TAIR10 Arabidopsis protein database. The Trinotate (version 1.0) pipeline was used to further annotate sequences (Grabherr, 2011). A total of 30,034 sequences (70%) had GO terms map (release 2014-10-16) (Ashburner et al., 2000), and 23,000 sequences (54%) had InterPro scan (version5-44.0) hits (Quevillon et al., 2005). SignalP (Petersen, Brunak, von Heijne, & Nielsen, 2011) detected 5,710 signal peptide cleavage sites and TmHMM (Krogh, Larsson, von Heijne, & Sonnhammer, 2001) detected 18,197 potential transmembrane domains amongst all the sequences. RepeatMasker was used to identify repetitive sequences in the transcriptome (Table 3.6). Although most TEs are transcriptionally silent (Lisch & Bennetzen, 2011), a total of 8,367 retroelements were identified in the transcriptome, the majority of which are LTR elements, covering 1.92% of the bases in the reference transcriptome. A total of 3,061 DNA transposons were also identified, covering 0.23% of the bases in the reference transcriptome. The number of retroelements and DNA transposons detected

across time was relatively stable and did not show any appreciable increase or decrease across time. It is likely that our findings are conservative, as many transcribed retroelements show low expression levels (F. Jiang, Yang, Guo, Wang, & Kang, 2012) and thus may not have been included in the reference transcriptome.

3.3.5 The Ceratopteris transcriptome is dynamic across early development

Unfortunately, it was determined that the samples grown in FM were contaminated with A_{CE}, therefore an analysis of the effects of $+A_{CE}$ versus $-A_{CE}$ treatment on gene expression across time was not possible. Nevertheless the data was useful for profiling the transcriptome of the male gametophyte across time. In order to identify genes with dynamic expression across male gametophyte development, a differential expression analysis was performed on pairs of consecutive time-points using the Wald test in DESeq2 with a 5% FDR and a biological-significance fold-change cutoff of 2. The differential expression analysis was performed on the 32,128 genes in the reference transcriptome. A design formula ("design =~ time + biological replicate") was specified to test the effect of time across samples, since the effects of condition are no longer being considered. Differential expression analyses were performed to find genes changing in expression between 0-3DAI, 3-3.5DAI, 3.5-4.5DAI, and 4.5-5.5DAI. A large number of differentially expressed genes were found, many of which have vast expression differences between time-points (Fig 3.3 and 3.4). As seen in Figure 3.3, between 0 and 3 days, 13,435 genes were differentially expressed, with 6,844 going up in expression and 6,591 going down in expression. The 0 day time-point captures genes stored in the dry spore and between these two time-points the spores are transitioning

from a dormant state to a metabolically active state. The differential expression analysis between 3 and 3.5 DAI found 2,253 differentially expressed genes (2,219 went up in expression and 34 went down in expression) and will likely capture genes that are involved in the changes that take place when the spore cracks open and becomes competent to respond to A_{CE}. The gametophyte will not develop as a male if it is not grown in the presence of A_{CE} from this point onward, thus the genes that encode gene products involved in the perception of A_{CE} and the initiation of downstream response should be present from day 3 or 3.5 onward (Banks et al., 1993b). Between 3.5-4.5 DAI 4,441 genes are differentially expressed; 3,537 increase in expression and 904 decrease in expression. The number of differentially expressed genes between 4.5 and 5.5 DAI is greater still, with 4,175 genes showing statistically significant differential expression, 3,116 of which are increasing in expression and 1,059 are decreasing. Table 3.7 lists the molecular function GO terms associated with the differentially expressed genes at each pair of time-points and Appendix B details the enriched GO terms amongst the differentially expressed genes. While it cannot be ruled out that the absolute expression of genes is changing, the large number of differentially expressed genes and the many GO terms are represented amongst these genes ultimately suggest that the male gametophyte transcriptome is dynamic.

3.3.6 A vast number of transcripts are stored in dormant spores

Our results show that a total of 17,280 genes are expressed across all the time-points assayed. Overall 18,437 genes are expressed in the dry spore alone. A total of 22,148 genes were expressed at 3 DAI, 22,086 were expressed at 3.5 DAI, 22,964 were

expressed at 4.5 DAI, and 24,459 were expressed at 5.5 DAI (Figure 3.5). It is noteworthy that there are so many stored transcripts in dry spores, which are dormant and metabolically inactive (Banks et al., 1993b). Our results are in agreement with previous estimates of gene expression in dry spores in which it was estimated that over 14,000 genes were expressed in spores (Salmi et al., 2005). GO terms were assigned to the dry spore transcripts and not surprisingly, as shown in Figure 3.6, there are a large number of GO terms associated with these transcripts. The GO terms mapped to the highest number of sequences are involved in metabolic processes; the most prevalent are macromolecular metabolic process, organic cyclic metabolic process, heterocycle metabolic process, cellular aromatic compound metabolic process, cellular nitrogen compound biosynthetic process, and nucleobase-containing metabolic process. These functional categories account for over 34% (17,478 sequences have these terms mapped to them) of the biological process GO terms mapped and also account for the majority of the functional categories present in gametophytes 5.5DAI (Figure 3.7). Overall, few differences were observed between the GO terms identified between the dry spore samples (Figure 3.6) and the gametophyte samples at 5.5DAI (Figure 3.7). Salmi *et al.* performed a study analyzing Ceratopteris ESTs early in development and similarly found that GO terms related to metabolism predominated in transcripts present in the spore (Salmi et al., 2005).

In both pollen and spores, translation is necessary for germination, evidenced by the fact that the inhibition of translation by cycloheximide blocks germination (Fernando, Owens, Yu, & Ekramoddoullah, 2001; Raghavan, 1970, 1971; Salmi et al., 2005). However, fern spores can successfully germinate in the absence of transcription. Raghavan *et. al* have shown that in two ferns closely related to Ceratopteris, *Asplenium*

nidus and *Pteridium aquilinum*, while transcription of mRNA is needed for elongation of the protonema, it is not needed for germination and initiation of the protonema (Raghavan, 1965, 1968; Raghavan & Tung, 1967). Since transcription is not necessary for germination, all the transcripts needed for germination of the spore and the formation of a rhizoid are pre-formed and stored in the spore (Raghavan, 1971). Protein synthesis from pre-formed mRNAs is needed for the gametophyte to elongate and form an independent, photosynthetic gametophyte (Raghavan, 1970). Hence it is not surprising that a large number (7,173 sequences) of the transcripts present in the dry Ceratopteris spore samples relate directly to the production of proteins.

3.3.7 Unsupervised clustering was performed to group genes based on temporal expression profiles

Due to the changes in the transcriptional landscape across time as evidenced by the number of genes changing between time-points and the large number of GO terms associated with these genes, patterns are difficult to see in the data. Unsupervised clustering is a useful way to aid in the identification and visualization of patterns. To cluster the data, fold changes and p-values from the differential expression analysis, adjusted for multiple testing, were used in determining whether a given gene maintained the same level of expression between two time-points, or displayed a statistically significant change in gene expression from one time-point to the next. From one time-point to the next, each gene was classified as going up in expression, down in expression, or not changing. Genes were then grouped based on common expression patterns. Patterns deemed particularly biologically interesting (genes increasing in expression upon

the addition of A_{CE}, genes decreasing in expression upon the addition of A_{CE}, and genes increasing in expression across time), as well as patterns encompassing a large number of genes (over 500) were graphed and an enrichment test was performed on each cluster (Fig 3.8).

The group of 71 genes that increase at each time-point (Fig. 3.8) mainly contains genes similar to genes involved in primary metabolism, many of which have established roles in plant growth and development. The enriched biological process GO terms in this group are carbohydrate metabolic process and metabolic process (Fig. 3.8 and Table 3.8). Thus it seems that transcripts encoding proteins involved in primary metabolism are increasing across time. This fits with what we know about Ceratopteris development, since during the time-points assayed the gametophyte is rapidly growing and cells are dividing and progressing from a metabolically inactive, dormant spore to a fully independent, photosynthesizing gametophyte (Banks et al., 1993a).

Given how dynamic the transcriptome is, the cluster with 9,981 genes that do not change across time could be very useful for time-course gene expression assays utilizing qRT-PCR. This cluster has a number of genes similar to house-keeping genes such as *ELONGATION FACTOR 1-α*, which was used as the reference gene for the qRT-PCR performed to validate this RNA-Seq data; these genes can be used as reference genes in the future. Also in this cluster are a couple of genes similar to those encoding products with GA-related functions. Both a gene similar to the gene encoding GA20ox, an enzyme involved in GA biosynthesis and genes similar to the gene encoding the GA receptor (GID1A) are in this group of genes (Table 3.8).

Another interesting cluster of genes seen in Figure 3.8 is the group of 4,806 genes that exhibit an initial decrease in expression between 0 DAI and 3 DAI and then do not change significantly from that point onward. These could be genes encoding products needed immediately upon germination and in the initial stages of early gametophyte development. The biological process GO terms enriched in this cluster include cell communication, protein metabolic process, and translation. No genes similar to GA responsive genes were found; however, this cluster of genes does contain genes similar ABA related genes. ABA is known to be involved in an array of processes, including seed dormancy in angiosperms although its role in ferns is thus far unknown. ABA and GA are antagonistic phytohormones which are together known to mediate the breakdown of seed dormancy (Xi, Liu, Hou, & Yu, 2010). The sequence comp108438_c1 is similar to *REGULATORY COMPONENTS OF ABA RECEPTOR 2* (PYL7/RCAR2), an ABA sensor (reviewed in (Sheard & Zheng, 2009)) and comp109917_c1 is similar to *ABSCISIC ACID INSENSITIVE 3* (*ABI3*), a transcription factor involved in the downstream responses to ABA (Table 3.8) (Nakashima et al., 2006).

3.3.8 Gene expression profiles of genes similar to GA-related genes

Due to the evidence suggesting that A_{CE} and GA have a common biosynthetic pathway in Ceratopteris, we are particularly interested in expression across time of genes similar to genes known to be involved in GA-related processes (Furber et al., 1989; Hickok, 1983; Takeno et al., 1989; Tanaka et al., 2014; T. R. Warne & Hickok, 1989; Yamane, 1998b; Yamane, Nohara, Takahashi, & Schraudolf, 1987b). A number of genes similar to those encoding gene products involved in GA-related processes were found in

the previous RNA-Seq experiment, many of which are differentially expressed between $-A_{CE}$ and $+A_{CE}$ conditions. The corresponding genes were identified using BLASTn in the time-course dataset and expression patterns are shown in Figure 3.9. While the functions of these genes are not known in Ceratopteris, they are excellent candidates for reverse genetics experiments to test the hypothesis that GA and A_{CE} share a common biosynthetic and signaling pathway.

Figure 3.8 A depicts the expression of genes similar to those involved in the GA signaling pathway hypothesized in Chapter 2 (Figure 2.6). To genes containing MYB transcription factors increase in expression across time and both were found to increase in expression in A_{CE} treated samples in Chapter 2. A gene with similarity to *MYB/120/33/101* and another with similarity to *MYB3R* in Arabidopsis were identified. The gene product of *MYB33* in Arabidopsis is a GAMYB and is a regulator of GA-related responses; it is de-repressed in the presence of GAs (Gocal et al., 2001). Interestingly, both of the genes with MYB domains increase in expression after the point in which A_{CE} was added to media (beginning 3.5DAI) and continues to increase at each subsequent time-point. It is plausible that one or both of these genes are de-repressed in the presence of GAs, including possibly A_{CE}, and could be involved in regulation of GA responses and possibly of sex determination. *GID1A* encodes the GA receptor in Arabidopsis, *GAI* encodes a DELLA domain transcription factor, which represses GA responses; in Arabidopsis GA binding to GID ultimately leads to the degradation of DELLA (Sun, 2011). Neither of the genes similar to *GID1A* change drastically in expression between time-points. Although *GID1Aa* appears to decrease in expression across time, this decrease was not found to be statistically significant. These expression

patterns are similar to those observed in *GID1* homologs in developing Lygodium gametophytes (Tanaka et al., 2014).

In Figure 3.9 B., the expression of genes similar to genes involved in GA biosynthesis are shown. *ENT-COPALYL DIPHOSPHATE SYNTHASE*/*ENT-KAURENE SYNTHASE* (*CPS/KS*), *GA 20-OXIDASE* (*GA20ox*), *ENT-KAURENE OXIDASE* (*KO*), and *GA 3-OXIDASE* (*GA3ox*) encode key enzymes in the Arabidopsis GA biosynthetic pathway (Sun & Kamiya, 1994). *CPS/KS* was found to be differentially expressed in the RNA-Seq data set discussed in Chapter 2, exhibiting higher levels of expression in $-A_{CE}$ samples. Other than the gene with similarity to KO, which exhibits somewhat high expression that does not change significantly between time-points, the remainder of the potential GA biosynthesis genes maintain relatively low levels of expression, increasing only between 4.5 and 5.5 DAI. Tanaka *et al.* found that *Lj_CPS/KS*, *Lj_KO*, and *Lj_GA20ox* were preferentially expressed in mature gametophytes that secrete antheridiogen, and showed much higher expression (10-20 times greater) than in young gametophytes that do not secrete antheridiogen (Tanaka et al., 2014). Tanaka *et al.* also found that the levels of *Lj_GID1* and *Lj_GA3ox* were expressed higher in young immature prothalli then in mature gametophytes. Thus, a split model of antheridiogen biosynthesis in Lygodium was proposed, with early-maturing gametophytes expressing GA biosynthetic genes with the exception of *GA3ox*. In the model, it is proposed that GA biosynthetic genes are used to produce antheridiogen, which is then excreted into the environment and taken up by later maturing gametophytes, which express GA3ox and thus modify antheridiogen into a bioactive GA. However the data in Figure 3.9 B shows that a gene similar to a gene encoding GA3ox maintains a stable level of expression until

4.5 DAI, after which point it exhibits a statistically significant increase in expression and furthermore a gene similar to *GA3ox* is not one of the genes found to be differentially expressed between gametophytes treated with $+A_{CE}$ and $-A_{CE}$ in Chapter 2. Thus the data in Figure 3.9 B along with the differential expression analysis in Chapter 2 suggest that Ceratopteris likely does not have a split antheridiogen biosynthetic pathway such as the one seen in Lygodium. This is not surprising because neither GA_{73} nor GA_{73} methyl ester, which are the antheridiogens in Lygodium substitute for the antheridiogen of Ceratopteris (unpublished observation). It is possible that *GID1A* expression and *GA3ox* expression could significantly drop at times past 5.5DAI or increase in hermaphrodites that secrete A_{CE}.

Figure 3.9 C. shows the temporal expression profiles of genes with similarity to transcription factors known to be involved in downstream GA responses in Arabidopsis. In Arabidopsis, the genes *LOM*, *LRP*, *MOTHER OF FT AND TFL1* (*MFT*), *SCL*, and a GRAS family gene member all encode transcription factors involved in GA responses in angiosperms (Gou et al., 2010; Xi et al., 2010) and genes similar to these were found to be differentially expressed between $+A_{CE}$ and $-A_{CE}$ treatments in Chapter 2. Figure 3.9 C shows an increase in expression of all of these transcription factors across time. *LOM* appears to have the largest increase in expression across time, however only the change in expression between 0 and 3 days was statistically significant. The expression profile of LRP shows the lowest level of expression and exhibits the smallest increase in expression across, though the increase between 3 and 3.5 days as well as the increase between 4.5 and 5.5 days is statistically significant. It is possible that these genes are not only slightly increasing in expression in males across time, but are also decreasing in

expression in hermaphrodites, which may have led to the genes being detected as differentially expressed in the RNA-Seq experiment discussed in Chapter 2.

3.3.9 RNA-Seq expression analysis results were validated by qRT-PCR

In order to assess the validity of the RNA-Seq data and expression analysis results, qRT-PCR was performed to assess the relative expression of ten genes between 3 time-points (3.5, 4.5, and 5.5 DAI). Melting curve analysis showed individual peaks for each target, and thus only a single target was amplified by each set of primers. Ten genes were assayed in order to calculate relative expression between 3.5-4.5 DAI, and the same ten genes were assayed in order to calculate relative expression between 4.5-5.5 DAI. As shown in Figure 3.10, the results of the qRT-PCR and RNA-Seq expression analysis agree 90% of the time.

3.4 Conclusion

In conclusion, a time-course RNA-Seq experiment was performed on young male Ceratopteris gametophytes. A transcriptome was assembled and a differential expression analysis was performed on each pair of time-points. As a result of these analyses several conclusions have become clear. First, this experiment shows that the transcriptome of gametophytes early in development is dynamic, involving changes in the expression of many genes. It is now clear that dynamic changes in transcript abundance and complexity occur during early male gametophyte development. It will be interesting to investigate to what extent these changes are due to the differential decay of stored transcripts and/or the synthesis of new transcripts.

These experiments have also shown that many more genes are expressed in the Ceratopteris male gametophyte than in the Arabidopsis gametophyte. In Arabidopsis, the number of genes expressed in the male gametophyte (pollen) is estimated to be ~4,172 (Loraine et al., 2013). It has been previously hypothesized that the large number of transcripts expressed in Ceratopteris gametophytes may be due to the fact that the Ceratopteris gametophytes are independent of the sporophyte and are also morphologically more complex than the pollen gametophyte(Salmi et al., 2005). The large number of genes expressed early in development in the Ceratotperis gametophyte, as well as the array of molecular function GO terms observed in the transcriptome and amongst the genes with dynamic expression across time fits with this hypothesis.

Additionally, the RNA-Seq experiment discussed here underscores that although the dry spore is dormant, a number of transcripts are stored, poising the spore for germination and differentiation. There were 18,437 genes present in the spore, representing a gamut of biological processes. It has been known for years that spores need *de novo* protein synthesis to germinate (Raghavan, 1965, 1968; Raghavan & Tung, 1967).

Lastly, the results of this study indicate that Ceratopteris does not exhibit the split antheridiogen biosynthetic pathway that Lygodium is proposed to utilize. However it is plausible that Ceratopteris, Lygodium, and Arabidopsis share many of the same GA biosynthesis and signal transduction components. RNAi knock-down experiments can be utilized to test the hypothesis that the signal transduction and biosynthesis components involved in GA signal transduction in Arabidopsis are also at work in Ceratopteris.

Table 3.1. Experimental Design of the time-course experiment, taking into consideration the loss of one replicate due to poor sequence quality. Each "X" represents one biological replicate. A_{CE} was added to the samples after harvesting on day 3. Poor sequence quality was observed for one sample at 0 days, and thus this sample was excluded from all further analyses; 3 biological replicates were obtained for all other conditions assayed.

	0days	**3days**	**3.5days**	**4.5days**	**5.5days**
$-A_{CE}$	**x x**	**x x x**	**x x x**	**x x x**	**x x x**
$+A_{CE}$			**x x x**	**x x x**	**x x x**

Table 3.2. Number of reads for each sample used in the transcriptome assembly. The number of reads shown here represents the reads which passed cleaning and quality control.

Original File	Number of Reads in Original File
0 DAI sample 1, left reads	95239080
0 DAI sample 1, right reads	95285553
0 DAI sample 2, left reads	72774222
0 DAI sample 2, right reads	72900376
3 DAI sample 1, left reads	66608806
3 DAI sample 1, right reads	66600206
3 DAI sample 2, left reads	59127753
3 DAI sample 2, right reads	59116028
3 DAI sample 3, left reads	86263240
3 DAI sample 3, right reads	86261806
$+A_{CE}$ 3.5 DAI sample 1, left reads	62647587
$+A_{CE}$ 3.5 DAI sample 1, right reads	62645607
$+A_{CE}$ 3.5 DAI sample 2, left reads	79717595
$+A_{CE}$ 3.5 DAI sample 2, right reads	79705624
$+A_{CE}$ 3.5 DAI sample 3, left reads	73997252
$+A_{CE}$ 3.5 DAI sample 3, right reads	73993591
$-A_{CE}$ 3.5 DAI sample 1, left reads	79194083
$-A_{CE}$ 3.5 DAI sample 1, right reads	79215408
$-A_{CE}$ 3.5 DAI sample 2, left reads	71802196
$-A_{CE}$ 3.5 DAI sample 2, right reads	71805028
$-A_{CE}$ 3.5 DAI sample 3, left reads	79031911
$-A_{CE}$ 3.5 DAI sample 3, right reads	79041270
$+A_{CE}$ 4.5 DAI sample 1, left reads	44025358
$+A_{CE}$ 4.5 DAI sample 1, right reads	44024430
$+A_{CE}$ 4.5 DAI sample 2, left reads	69494275
$+A_{CE}$ 4.5 DAI sample 2, right reads	69513890
$+A_{CE}$ 4.5 DAI sample 3, left reads	63935871
$+A_{CE}$ 4.5 DAI sample 3, right reads	63932643
$-A_{CE}$ 4.5 DAI sample 1, left reads	69920326
$-A_{CE}$ 4.5 DAI sample 1, right reads	69937438
$-A_{CE}$ 4.5 DAI sample 2, left reads	65432261
$-A_{CE}$ 4.5 DAI sample 2, right reads	65482517
$-A_{CE}$ 4.5 DAI sample 3, left reads	60009087
$-A_{CE}$ 4.5 DAI sample 3, right reads	60011590
$+A_{CE}$ 5.5 DAI sample 1, left reads	65291536
$+A_{CE}$ 5.5 DAI sample 1, right reads	65293809
$+A_{CE}$ 5.5 DAI sample 2, left reads	71392835
$+A_{CE}$ 5.5 DAI sample 2, right reads	71402223

Table 3.2 Continued

+A_{CE} 5.5 DAI sample 3, left reads	53856128
+A_{CE} 5.5 DAI sample 3, right reads	53860524
-A_{CE} 5.5 DAI sample 1, left reads	72927063
-A_{CE} 5.5 DAI sample 1, right reads	72964371
-A_{CE} 5.5 DAI sample 2, left reads	52995868
-A_{CE} 5.5 DAI sample 2, right reads	53055277
-A_{CE} 5.5 DAI sample 3, left reads	84545075
-A_{CE} 5.5 DAI sample 3, right reads	84550980

Table 3.3. Summary of the data cleaning input/output. The table includes programs used in data cleaning, input number of reads, output number of reads, and the percentage of the original starting number of reads remaining. Of the original ~3.4 billion reads obtained from the sequencing facility, 94% remained by the end of the entire cleaning workflow.

Program	Input No. Reads	Output No. Reads	Percent Remaining
Trimmomatic	3,398,072,444	3,348,264,002	99%
DeconSeq	3,348,264,002	3,218,137,063	95%
clean_adapter.pl	3,218,137,063	3,200,829,597	94%
getpairs.pl	3,200,829,597	3,190,064,796	94%

Table 3.4. Assembly statistics for the full transcriptome assembly and for the reference transcriptome assembly. In running Trinity, the min_contig was set to 200 (Grabherr et al., 2011a; B. Li & Dewey, 2011) and therefore it is not surprising that the minimum sequence length was 201bp. Also not surprisingly, most of the sequences that were removed from the assembly when making the reference were short sequences. The vast majority of these were removed when filtering out lowly expressed transcripts, thus many of them were likely spurious assemblies.

Assembly Statistics	**Full Assembly**	**Reference Assembly**
Number predicted transcripts	395,694	42,798
Sum length	282,019,132	105,320,862
N50	1,170	3,062
Min length	201	201
Max length	17,316	17,316
Average length	713	2,460
Median length	376	2,174

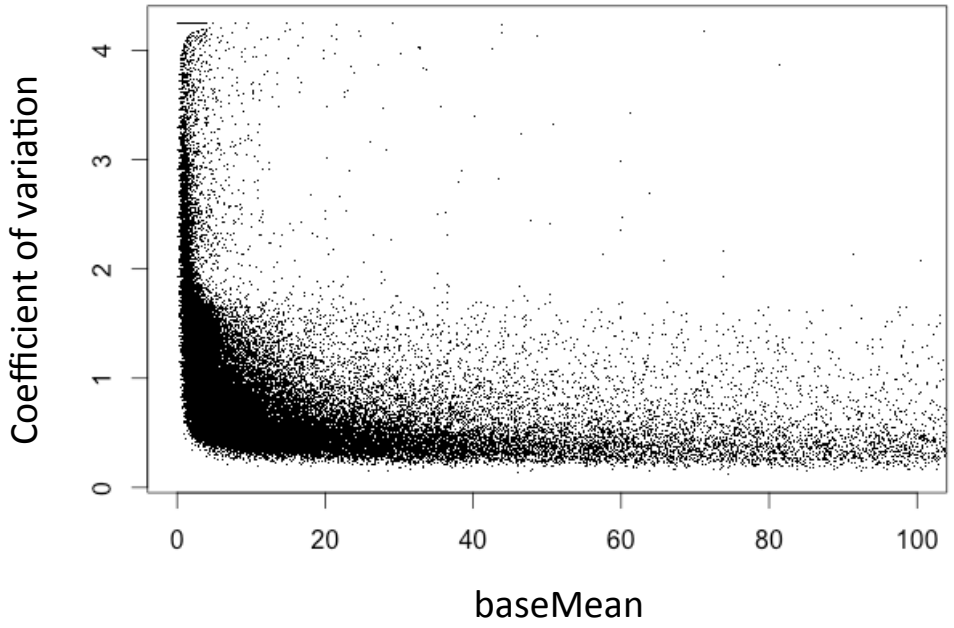


Figure 3.1. The baseMean versus the coefficient of variation for all genes with read support. The baseMean, shown on the x-axis, is the mean of counts, normalized based on library size. The coefficient of variation, shown on the y-axis, is the ratio of the standard deviation to the mean. As expected, at low counts the expression is variable and thus the coefficients of variation are larger and more varied.

Table 3.5. Table detailing the creation of a reference assembly.

	Starting number of sequences	Number of sequences removed	Number of sequences remaining
After CPM filtering	395,694	328,769	66,925
After annotation filtering	66,925	24,127	42,798

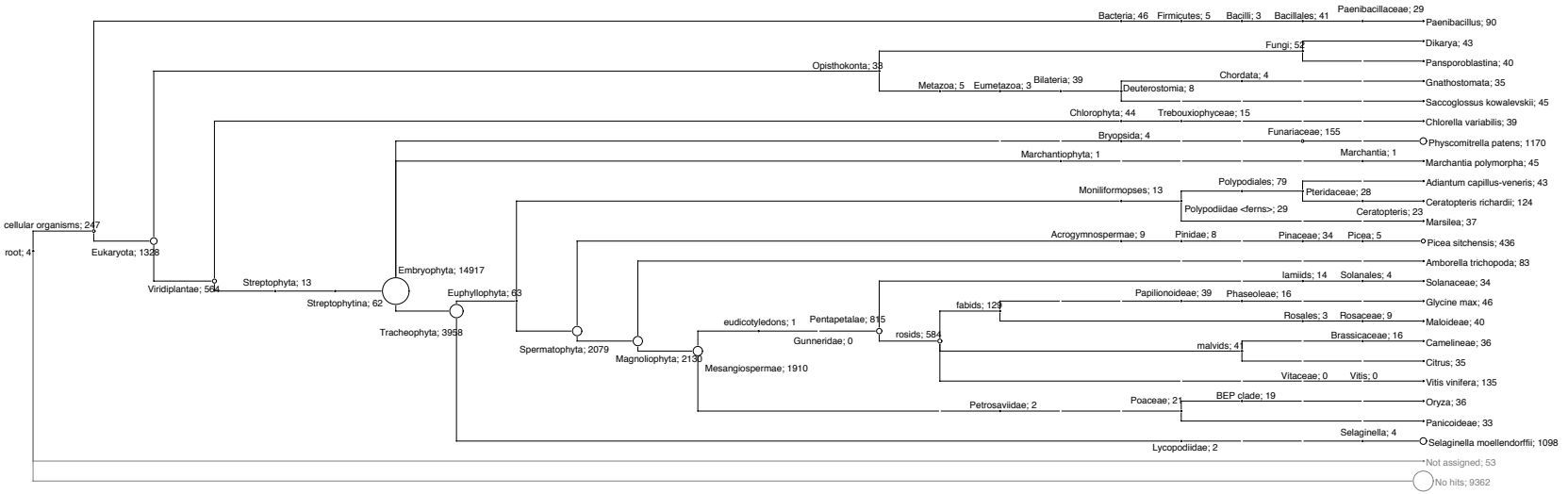


Figure 3.2. Cladogram of the taxonomic distribution of sequences in the reference transcriptome.

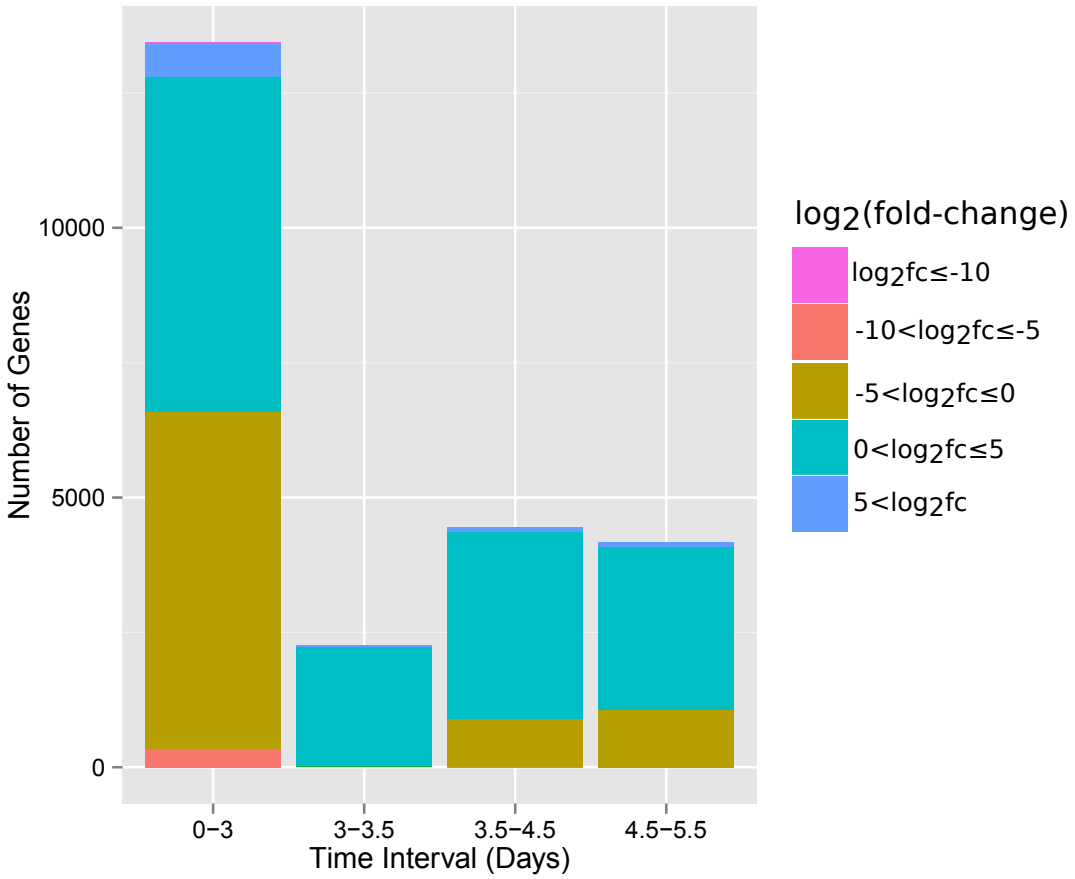


Figure 3.3. Number of differentially expressed genes between pairs of consecutive time-points. A histogram shows the number of differentially expressed genes along the y-axis and the time-intervals along the x-axis. The bars of the histogram are colored according to the $\log_2$(fold-change) of the differentially expressed genes. Only two genes exhibited a $\log_2$(fold-change) < -10, both during the 0-3DAI interval.

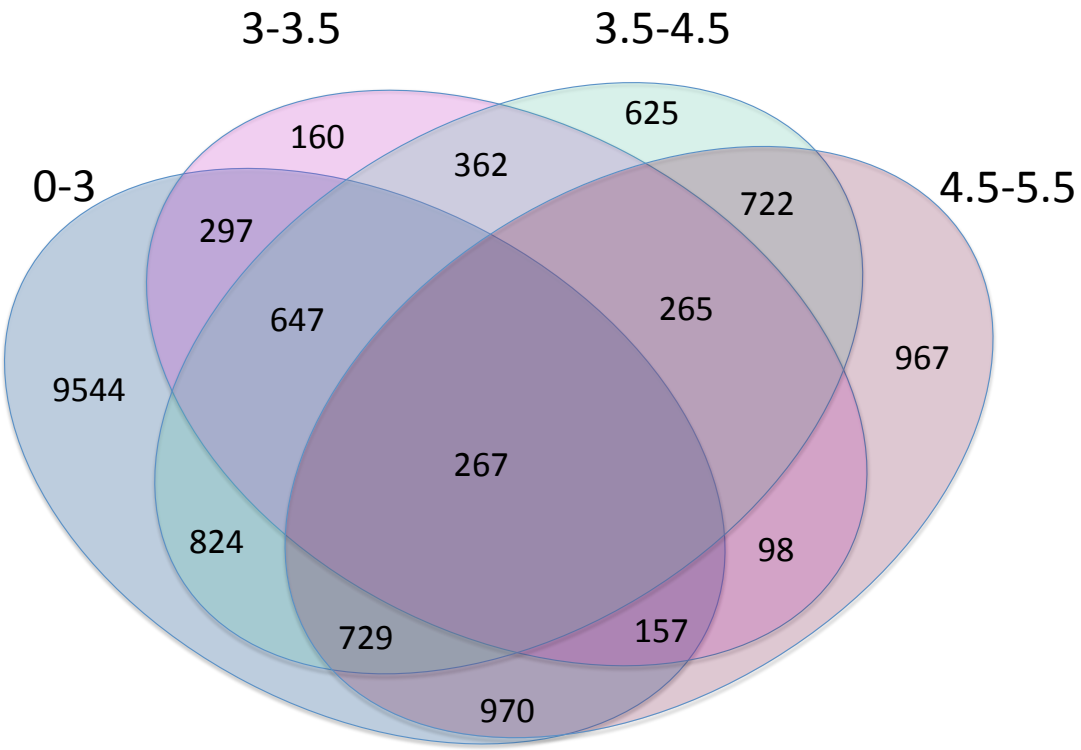


Figure 3.4. Venn diagram of differentially expressed genes between each pair of time-points.

Table 3.6. Transposable elements identified in the reference transcriptome. RepeatMasker was used to identify transposon sequences in the Ceratopteris transcriptome. A number of transposons were identified, including a significant number of various classes of Retroelements and DNA transposons.

Transposon class	Number of elements	Length occupied (bp)	Percentage of bases covered
Retroelements	8367	2,614,173	1.92 %
SINEs	12	745	0.00%
Penelope	1	129	0.00%
LINEs	579	41699	0.03%
R2/R4/NeSL	2	99	0.00%
RTE/Bov-B	17	920	0.00%
L1/CIN4	510	37495	0.03%
LTR elements	7776	2571729	1.89%
Ty1/Copia	4270	1179299	0.87%
Gypsy/DIRS1	3449	1385686	1.02%
DNA transposons	3061	315980	0.23%
hobo-Activator	1114	147684	0.11%
Tc1-IS630-Pogo	16	984	0.00%
Tourist/Harbinger	143	15887	0.01%
Other	3	141	0.00%

Table 3.7. Molecular function GO terms of differentially expressed genes between consecutive time-points. A number of molecular function GO terms are observed in sets of differentially expressed genes. Only GO term categories containing at least 50 sequences are included.

GO term	Number of Sequences
In DEGs between 0 and 3 DAI	
2-alkenal reductase [NAD(P)] activity	56
aminoacyl-tRNA ligase activity	54
antioxidant activity	60
ATP binding	1065
ATP-dependent helicase activity	67
calcium ion binding	119
carboxy-lyase activity	52
carboxylic ester hydrolase activity	58
cation-transporting ATPase activity	68
channel activity	50
chromatin binding	71
copper ion binding	150
cytoskeletal protein binding	72
disulfide oxidoreductase activity	59
electron carrier activity	126
endopeptidase activity	104
enzyme binding	87
enzyme regulator activity	101
flavin adenine dinucleotide binding	65
GTP binding	203
GTPase activity	132
heme binding	59
hydro-lyase activity	56
hydrogen ion transmembrane transporter activity	77
hydrolase activity, acting on carbon-nitrogen (but not peptide) bonds	57
hydrolase activity, hydrolyzing O-glycosyl compounds	105
identical protein binding	78
iron ion binding	89
iron-sulfur cluster binding	86
isomerase activity	199
magnesium ion binding	84
metal ion transmembrane transporter activity	111
metallopeptidase activity	52

Table 3.6 Continued

monooxygenase activity	66
NAD binding	57
NADP binding	54
nuclease activity	62
nucleotidyltransferase activity	115
organic anion transmembrane transporter activity	85
oxidoreductase activity, acting on NAD(P)H	74
oxidoreductase activity, acting on paired donors, with incorporation or reduction of molecular oxygen	105
oxidoreductase activity, acting on the aldehyde or oxo group of donors	60
oxidoreductase activity, acting on the CH-OH group of donors, NAD or NADP as acceptor	154
phosphate transmembrane transporter activity	55
phosphoprotein phosphatase activity	109
protein complex binding	54
protein heterodimerization activity	106
protein serine/threonine kinase activity	386
protein transporter activity	69
pyridoxal phosphate binding	70
S-adenosylmethionine-dependent methyltransferase activity	51
secondary active transmembrane transporter activity	101
sequence-specific DNA binding	86
sequence-specific DNA binding transcription factor activity	210
serine-type peptidase activity	57
signal transducer activity	107
structural constituent of ribosome	312
transferase activity, transferring acyl groups other than amino-acyl groups	113
translation elongation factor activity	53
translation initiation factor activity	54
ubiquitin-protein ligase activity	62
UDP-glucosyltransferase activity	64
unfolded protein binding	94
zinc ion binding	326
In DEGs between 3 and 3.5 DAI	
tetrapyrrole binding	0
hydrolase activity, hydrolyzing O-glycosyl compounds	63
protein heterodimerization activity	52

Table 3.6 Continued

copper ion binding	53
sequence-specific DNA binding	52
phosphatase activity	63
UDP-glycosyltransferase activity	64
sequence-specific DNA binding transcription factor activity	130
RNA binding	71
ATP-dependent DNA helicase activity	62
ATP binding	422
protein serine/threonine kinase activity	164
mismatched DNA binding	69
inorganic cation transmembrane transporter activity	57
structural molecule activity	52
transferase activity, transferring hexosyl groups	96
P-P-bond-hydrolysis-driven transmembrane transporter activity	62
anion transmembrane transporter activity	55
isomerase activity	72
calcium ion binding	50
transferase activity, transferring acyl groups	56
coenzyme binding	81
oxidoreductase activity, acting on the CH-OH group of donors, NAD or NADP as acceptor	56
zinc ion binding	124
lyase activity	79
ligase activity	63
methyltransferase activity	60
GTP binding	62
peptidase activity	63
In DEGs between 3.5 and 4.5 DAI	
tetrapyrrole binding	81
identical protein binding	57
transferase activity, transferring acyl groups other than amino-acyl groups	61
metal ion transmembrane transporter activity	95
flavin adenine dinucleotide binding	53
hydrolase activity, hydrolyzing O-glycosyl compounds	94
translation factor activity, nucleic acid binding	57
endopeptidase activity	52
sequence-specific DNA binding transcription factor activity	185
sequence-specific DNA binding	66

Table 3.6 Continued

GTP binding	121
helicase activity	60
NAD binding	50
signal transducer activity	84
copper ion binding	101
phosphoprotein phosphatase activity	70
oxidoreductase activity, acting on the aldehyde or oxo group of donors	58
carbon-carbon lyase activity	70
structural constituent of ribosome	72
acid-amino acid ligase activity	61
iron-sulfur cluster binding	50
UDP-glucosyltransferase activity	54
chromatin binding	51
electron carrier activity	76
monooxygenase activity	64
iron ion binding	61
calcium ion binding	99
oxidoreductase activity, acting on the CH-CH group of donors, NAD or NADP as acceptor	63
ATP binding	795
magnesium ion binding	56
organic anion transmembrane transporter activity	54
hydrogen ion transmembrane transporter activity	57
oxidoreductase activity, acting on paired donors, with incorporation or reduction of molecular oxygen	82
nucleotidyltransferase activity	63
oxidoreductase activity, acting on the CH-OH group of donors, NAD or NADP as acceptor	106
secondary active transmembrane transporter activity	81
carbon-oxygen lyase activity	51
enzyme regulator activity	54
unfolded protein binding	52
enzyme binding	63
isomerase activity	141
zinc ion binding	210
methyltransferase activity	105
protein heterodimerization activity	71
protein serine/threonine kinase activity	296
GTPase activity	88

Table 3.6 Continued

mismatched DNA binding	73
cation-transporting ATPase activity	64
In DEGs between 4.5 and 5.5 DAI	
tetrapyrrole binding	86
identical protein binding	51
transferase activity, transferring acyl groups other than amino-acyl groups	75
monovalent inorganic cation transmembrane transporter activity	87
metal ion transmembrane transporter activity	93
flavin adenine dinucleotide binding	53
hydrolase activity, hydrolyzing O-glycosyl compounds	88
translation factor activity, nucleic acid binding	71
endopeptidase activity	66
sequence-specific DNA binding transcription factor activity	184
pyridoxal phosphate binding	51
oxidoreductase activity, acting on a sulfur group of donors	61
UDP-glycosyltransferase activity	63
sequence-specific DNA binding	67
GTP binding	133
signal transducer activity	76
copper ion binding	89
oxidoreductase activity, acting on the aldehyde or oxo group of donors	56
phosphoprotein phosphatase activity	70
carbon-carbon lyase activity	68
structural constituent of ribosome	81
ATPase activity, coupled to transmembrane movement of ions	62
acid-amino acid ligase activity	58
oxidoreductase activity, acting on NAD(P)H	57
iron-sulfur cluster binding	56
chromatin binding	55
cytoskeletal protein binding	52
electron carrier activity	91
monooxygenase activity	56
iron ion binding	56
ATP-dependent DNA helicase activity	66
calcium ion binding	85

Table 3.6 Continued

oxidoreductase activity, acting on the CH-CH group of donors, NAD or NADP as acceptor	59
ATP binding	735
magnesium ion binding	59
organic anion transmembrane transporter activity	68
oxidoreductase activity, acting on paired donors, with incorporation or reduction of molecular oxygen	74
nucleotidyltransferase activity	99
oxidoreductase activity, acting on the CH-OH group of donors, NAD or NADP as acceptor	107
secondary active transmembrane transporter activity	84
carbon-oxygen lyase activity	58
enzyme regulator activity	59
isomerase activity	151
zinc ion binding	212
methyltransferase activity	110
transferase activity, transferring hexosyl groups	111
protein heterodimerization activity	69
protein serine/threonine kinase activity	255
GTPase activity	86

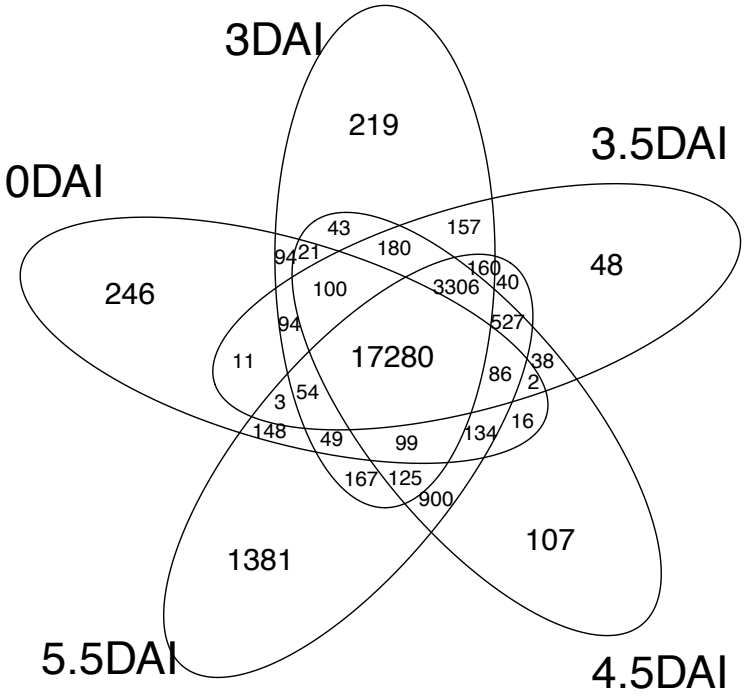


Figure 3.5. Venn diagram of genes expressed at each time-point. To be considered expressed genes must be expressed >0.3CPM. A total 17,280 genes are expressed in all five time-points assayed. A number of genes also show developmental stage-specific expression and the overall trend is that the number of genes expressed is increasing across time.

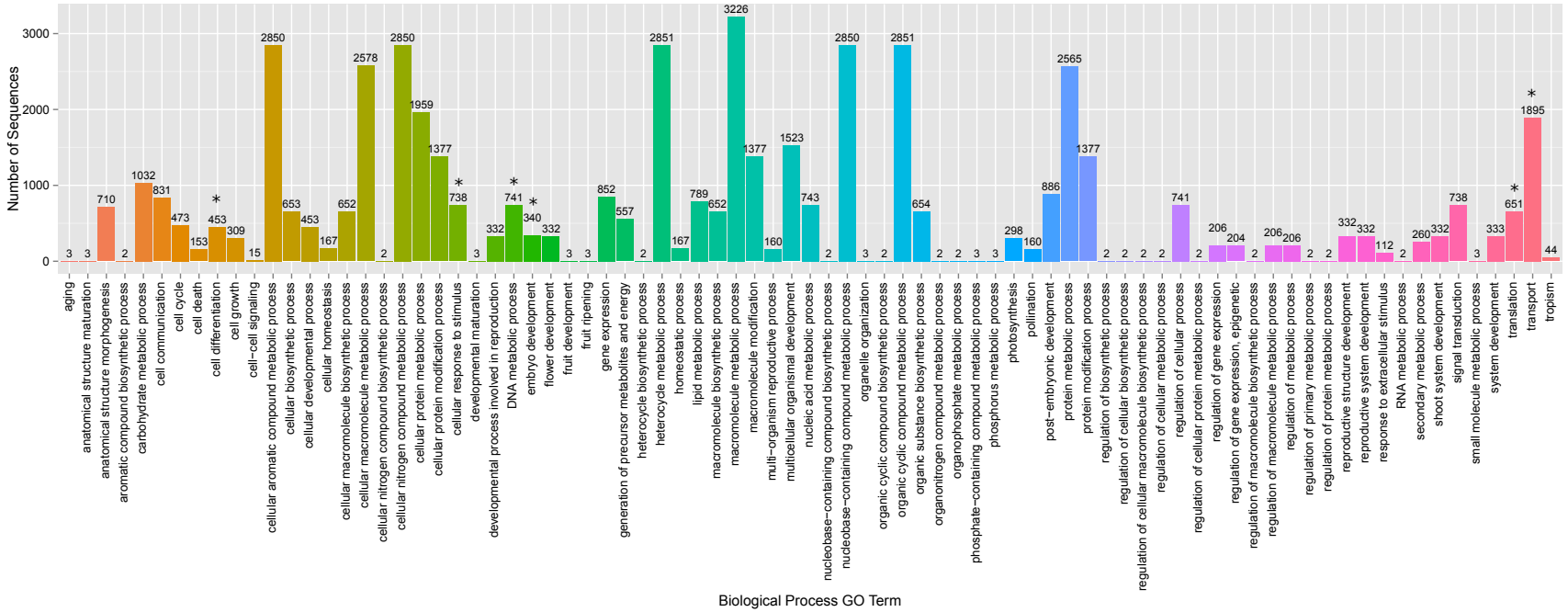


Figure 3.6. Biological process GOslim terms associated with transcripts present in dry spores. GO terms are listed on the x-axis and the number of sequences present in each GO term category is shown on the y-axis. The number of sequences present in each category is also listed above the bars. GO terms that were found to be enriched in the 0 DAI samples using a 5% FDR are starred.

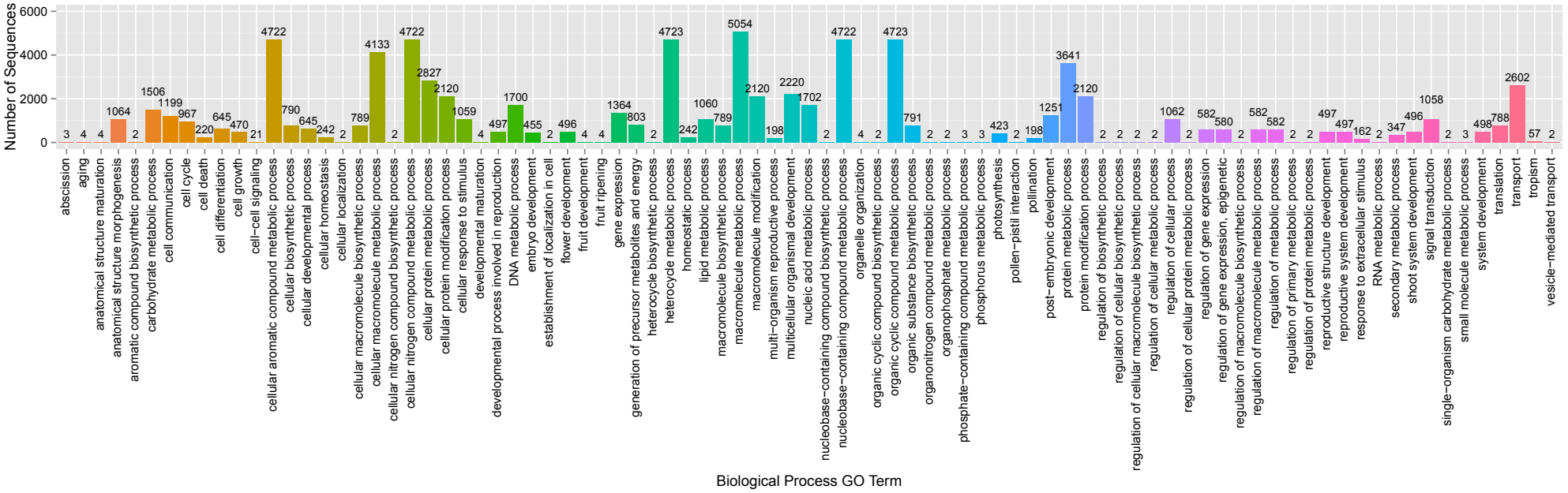


Figure 3.7. Biological process GOslim terms associated with transcripts present in gametophytes 5.5DAI. GO terms are listed on the x-axis and the number of sequences present in each GO term category is shown on the y-axis. The number of sequences present in each category is also listed above the bars.

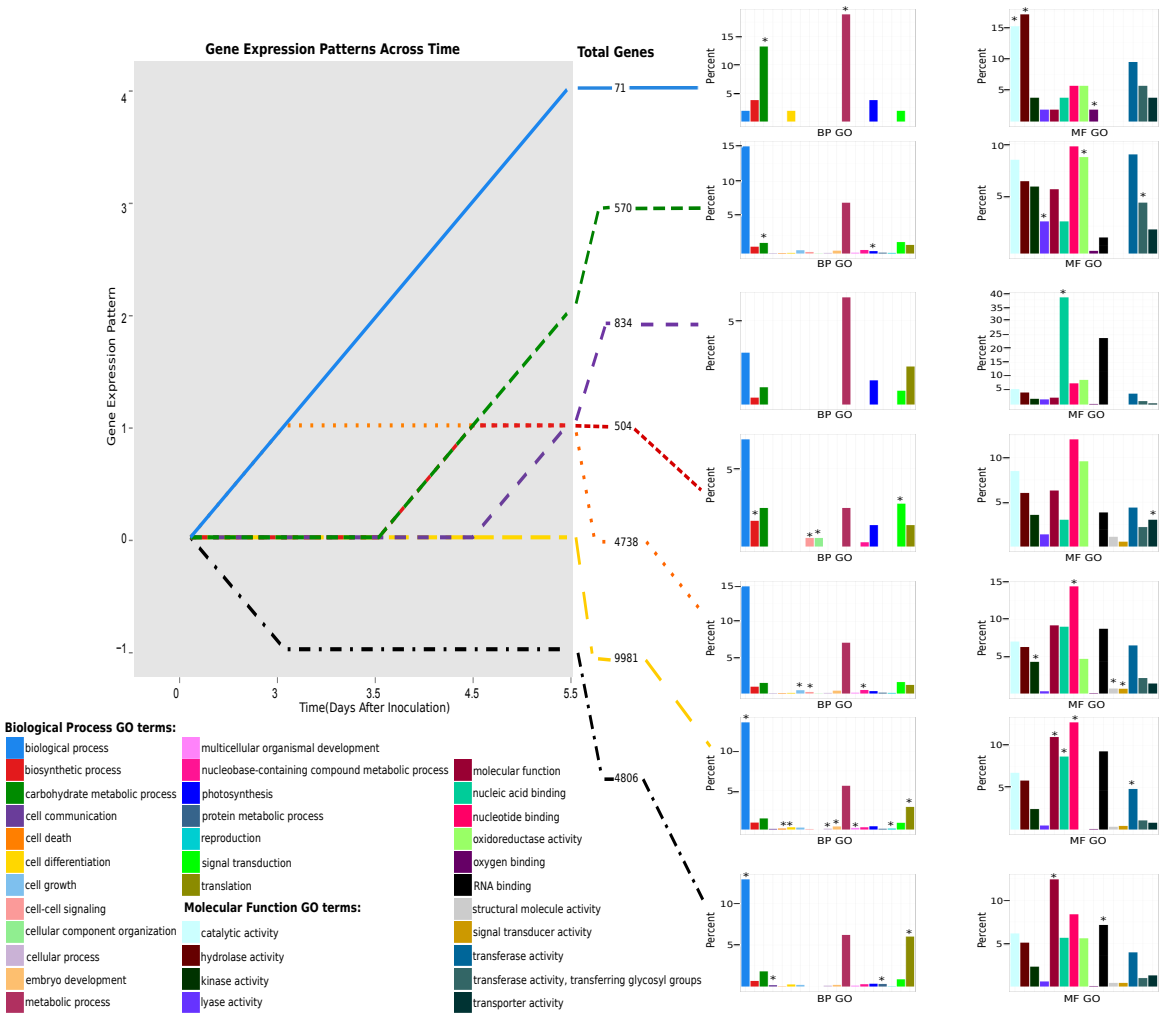
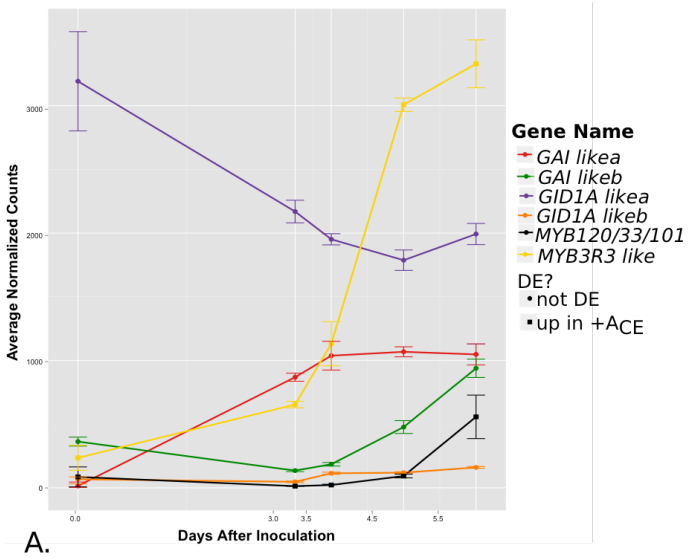


Figure 3.8. Patterns of select clusters of genes resulting from unsupervised clustering. The large graph shows the various time-points on the x-axis and the gene expression pattern on the y-axis. For the sake of readability, the patterns are shown using arbitrary y-axis values of 0 for no change, 1 for an, and -1 for a decrease in expression between time t and t+1. A different color line is shown for each cluster and to the right of the graph the total genes in each cluster is shown. To the far right are graphs for each cluster of genes. The graphs show the biological process (BP) and molecular function (MF) GO terms enriched in any of the clusters on the x-axes. The y-axes show the percentage of all sequences with GO terms mapped that are associated with that GO term. Biological process graphs are on the left and molecular function graphs are on the right. Starred bars indicate that the GO term is enriched in that specific cluster of genes.

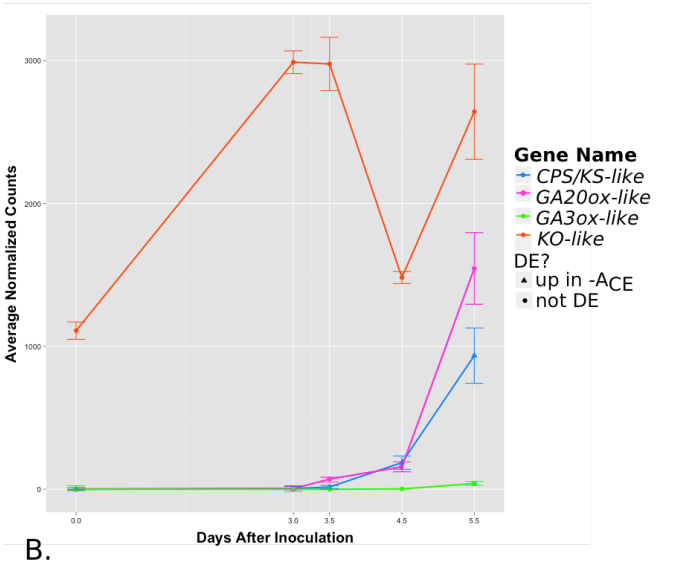
Table 3.8. List of *Ceratopteris* genes mentioned in the Chapter 3 discussion that are similar to *Arabidopsis* genes.

Ceratopteris gene number	**Most similar Arabidopsis gene**	**Arabidopsis Accession**	**BLASTx E-value**
71 gene cluster			
comp63305_c0_seq1	ADG2, APL1	AT5G19220.1	2.00E-76
comp63881_c0_seq1	GPAT5	AT3G11430.1	1.00E-84
comp118280_c0_seq1	GAPA-2	AT1G12900.1	E-170
comp115879_c0_seq1	PSAE-1	AT4G28750.1	2.00E-24
comp122757_c0_seq1	HCEF1	AT3G54050.2	E-160
comp113406_c0_seq1	LHCB5	AT4G10340.1	1.00E-94
4806 gene cluster			
comp108438_c1_seq2	PYL7, RCAR2	AT5G53160.1	2.00E-45
comp109917_c1_seq1	ABI3	AT3G24650.1	2.00E-39
9981 gene cluster			
comp109219_c5_seq1	ELF1A	AT5G60390.3/AT5G60390.1/AT1G07920.1/AT1G07930.1/AT1G07940.1	1.00E-46
comp111615_c0_seq1	ELF1A	AT5G60390.3/AT5G60390.1/AT1G07920.1/AT1G07930.1/AT1G07940.1	0
comp109998_c2_seq1	ELF1A	AT5G60390.3/AT5G60390.1/AT1G07920.1/AT1G07930.1/AT1G07940.1	1.00E-15
comp111599_c0_seq1	ELF1A	AT5G60390.3/AT5G60390.1/AT1G07920.1/AT1G07930.1/AT1G07940.1	0
comp120863_c0_seq1	ELF1A	AT5G60390.3/AT5G60390.1/AT1G07920.1/AT1G07930.1/AT1G07940.1	0.00E+00
GA signal transduction genes			
comp124120_c0	GAI a	AT1G14920.1	E-112
comp127127_c0	GAI b	AT1G14920.1	E-112
comp60529_c0	GID1A a	AT3G05120.1	E-110
comp103793_c0	GID1A b	AT3G05120.1	3.00E-93
comp59870_c0	MYB3R3	AT5G11510.2	3.00E-80

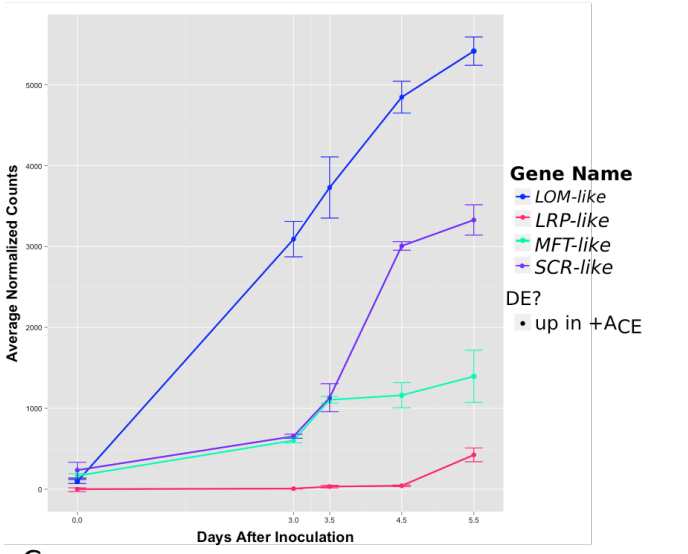
comp59469_c0	MYB 120/33/101 related	AT5G55020.1	1.00E-49
GA biosynthetic genes			
comp128084_c0	CPS/KS	AT4G02780.1/AT1G79460.1	E-138
comp125198_c0	GA20ox	AT4G25420.1	4.00E-87
comp118198_c0	KO	AT5G25900.1	E-118
comp161286_c0	GA3ox	AT4G21690.1	2.00E-50
GA related transcription factors			
comp108099_c0	LOM	AT3G60630.1	3.00E-48
comp59442_c1	LRP	AT5G66350.1	4.00E-28
comp122039_c0	MFT	AT1G18100.1	7.00E-46
comp124866_c0	SCL	AT5G66770.1	3.00E-77


Average Normalized Counts
3000
2000
1000
0
0.0
3.0
3.5
4.5
5.5
Days After Inoculation
Gene Name
GAI likea
GAI likeb
GID1A likea
GID1A likeb
MYB120/33/101
MYB3R3 like
DE?
not DE
up in +ACE

A.


Average Normalized Counts
3000
2000
1000
0
0.0
3.0
3.5
4.5
5.5
Days After Inoculation
Gene Name
CPS/KS-like
GA20ox-like
GA3ox-like
KO-like
DE?
up in -ACE
not DE

B.


Average Normalized Counts
5000
4000
3000
2000
1000
0
0.0
3.0
3.5
4.5
5.5
Days After Inoculation
Gene Name
LOM-like
LRP-like
MFT-like
SCR-like
DE?
up in +ACE

C.

Figure 3.9. Expression patterns of genes with BLASTx hits to proteins involved in GA-related processes. Days after inoculation is shown on the x-axes and the average normalized counts computed in DESeq2 is shown on the y-axes. The shape of the points depict whether or not the gene was found to be differentially expressed in the RNA-Seq experiment described in Chapter 2 (circle=not differentially expressed; triangle=up in -A_{CE}; square=up in +A_{CE}). A different colored line is shown for each gene and genes are referred to by the *Arabidopsis thaliana* abbreviations of the closest BLAST hit. A. Expression of genes with BLAST hits to proteins directly involved in the initial GA signal transduction pathway in Arabidopsis. B. Expression of genes with BLAST hits to proteins directly involved in the GA biosynthesis pathway in Arabidopsis. C. Expression of genes with BLAST hits to transcription factor products involved in the GA in Arabidopsis.

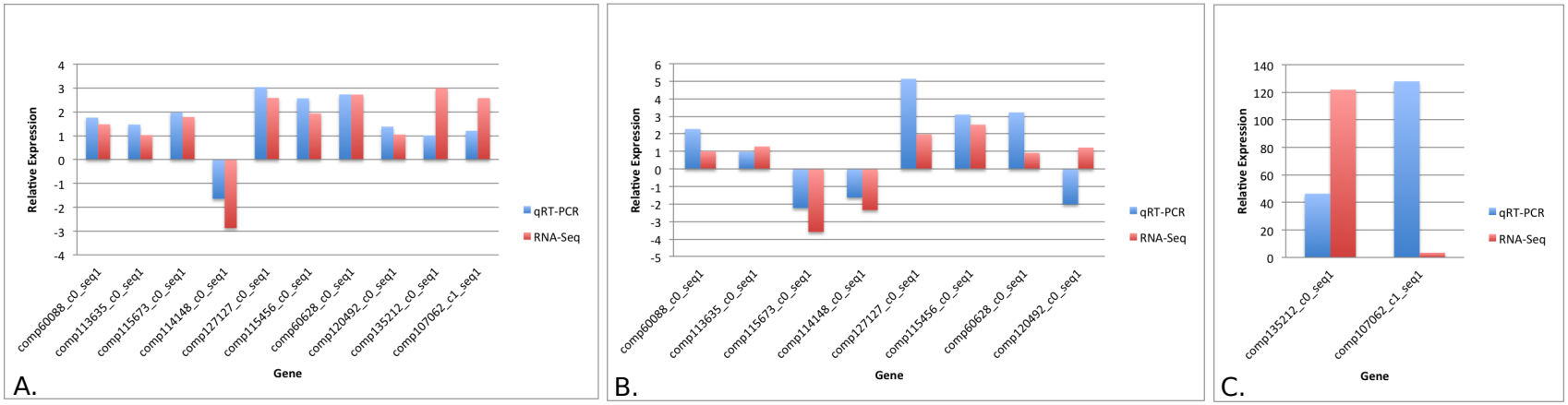


Figure 3.10. Results of the expression validation of RNA-Seq data using qRT-PCR. Relative expression is shown for ten genes between two pairs of time-points (3.5-4.5 DAI and 4.5-5.5 DAI). In 18/20 conditions, the qRT-PCR results (blue bars) agree with the RNA-Seq results (red bars). A. Relative expression of ten genes between the time-points 3.5DAI and 4.5 DAI. Genes with positive relative expression values were more highly expressed at 4.5 DAI than at 3.5 DAI. The qRT-PCR results validate the RNA-Seq results for 10/10 genes. To enhance readability the data for the relative expression of genes between 4.5 and 5.5 DAI was split between two graphs: B. shows the relative expression of eight genes with smaller relative expression values and C. shows the relative expression of two genes with large relative expression values. Genes with positive relative expression values were more highly expressed at 4.5 DAI than at 4.5 DAI. The qRT-PCR results validate the RNA-Seq results for 7/8 genes in B. and for 1/2 genes in C.

CHAPTER 4. CONCLUSION

Sex determination is a fundamental aspect of development, which allows generations of organisms to reproduce sexually. While sex is usually genetically determined, it can also be determined by environmental cues such as temperature and social environment (reviewed in (Atallah & Banks, 2015; Tanurdzic & Banks, 2004)). In Ceratopteris, the sex of the gametophyte, which is the haploid sexual phase of the land plant life cycle, is determined epigenetically by the social environment of the gametophyte. Sex is determined by the pheromone A_{CE}, which is emitted by hermaphrodite gametophytes upon loss of competence to respond to the male-inducing effects of A_{CE}. Thus, spores that develop in the absence of A_{CE} develop as hermaphrodites, while spores that germinate later, and in the presence of A_{CE}, develop as males (Banks, 1997a). While tests of epistasis between sex-determining mutants have been used to generate a genetic model of the sex determination pathway (Banks, 1994b, 1997d; Strain et al., 2001), these sex-determining genes have not been cloned. The molecular mechanisms involved in sex determination in Ceratopteris thus remains unsolved and, despite its significance in the survival of many species, little is known about the mechanisms involved in environmental sex determination.

The Ceratopteris genome is large and has not been sequenced, thus, cloning techniques are not feasible methods for cloning the sex determination genes.

The research presented here has used a different approach to find genes potentially involved in sex determination in Ceratopteris. Two RNA-Seq experiments were performed: one experiment allows comparison between gene expression levels in male ($+A_{CE}$) versus hermaphrodite ($-A_{CE}$) samples at 4.5 DAI and another RNA-Seq experiment details gene expression across time throughout early development in male ($+A_{CE}$) samples.

The goal of the initial RNA-Seq experiment described in Chapter 2 was to assemble a transcriptome, to identify differentially expressed genes between $\pm A_{CE}$ conditions, and to generate testable hypotheses for how A_{CE} controls the sex of the gametophyte at the gene expression level. A *de novo* transcriptome assembly was successfully performed using ~395 million 100bp paired-end reads, generating a transcriptome of gametophytes grown in the absence or presence of A_{CE}. Of the 82,820 predicted genes assembled, 1,163 are differentially expressed between $+A_{CE}$ and $-A_{CE}$ conditions. Overall, 89% of the differentially expressed genes are up-regulated in $+A_{CE}$ samples whereas only 11% are up-regulated in $-A_{CE}$ samples. Amongst the differentially expressed genes, a large number of genes similar to those involved in RNA processing and small RNA biogenesis are up-regulated by A_{CE}. Additionally a number of genes similar to those involved in histone modification, chromatin remodeling, and DNA methylation were identified in the genes up-regulated in $+A_{CE}$ samples. These results suggest that post-transcriptional regulation via RNAi and RNA processing, as well as large-scale reprogramming of the genome may be occurring after exposure to A_{CE}. The differential expression analysis also identified genes similar to those involved in GA signaling or response in Arabidopsis. This experiment led to the generation of an easily

testable model for how A_{CE} may be determining sex at a genetic and molecular level, which is currently being tested by RNAi.

The second RNA-Seq study provided gene expression data of male Ceratopteris gametophytes grown across early development. Time-points were chosen based on important developmental events: 0 DAI, 3 DAI, 3.5 DAI, 4.5 DAI, and 5.5 DAI and. A reference transcriptome was made and consists of 42,798 predicted transcripts. This reference was used in the differential expression analysis in order to identify genes that were differentially expressed between adjacent time-points. This experiment has shown that the transcriptome is dynamic across early gametophyte development: between 0-3 DAI 13,435 genes are differentially expressed, between 3-3.5 DAI 2,253 genes are differentially expressed, between 3.5-4.5 DAI 4,441 genes are differentially expressed, and between 4.5-5.5 DAI 4,175 genes are differentially expressed. The sequencing of the 0 DAI (dry spore) time-point has provided the first comprehensive look at the sequences of transcripts stored in the dry spore, at which point spores are poised in a dormant state, but contain all the transcripts needed to initiate germination and emergence of the prothallus (Raghavan, 1970, 1971, 1991; Raghavan & Tung, 1967). A total of 17,280 genes are expressed across all the time-points assayed and 18,437 genes are expressed in the dry spore at >0.3 CPM. Several conclusions can be framed based on the results of this time-course RNA-Seq experiment. First, the transcriptome of gametophytes early in development is dynamic, involving changes in the expression of the majority of genes detected. Second, the Ceratopteris male gametophyte has more transcripts present than the Arabidopsis gametophyte; it is possible that this is due to the fact that fern gametophytes are independent of the sporophyte and are morphologically more complex

than male gametophytes in angiosperms. Additionally, although the dry spore is dormant, a large number of transcripts are stored. There were numerous genes stored in the spore, representing a wide range of biological processes. The complexity of transcripts increases even more as gametophytes germinate and become metabolically and photosynthetically active. Finally, the results of this study also suggest that Ceratopteris does not exhibit the split antheridiogen biosynthetic pathway that is proposed to exist in Lygodium, another homosporous fern (Tanaka et al., 2014).

Overall, the RNA-Seq experiments described here provide the foundation for identification of the sex determination genes in Ceratopteris. These experiments have also provided insight into gene expression profiles of developing gametophytes. Additionally, as a result of these studies, Ceratopteris now has publically available high quality transcriptomics data. These transcriptome sequences provide a valuable resource for other researchers and could lead to the acceleration of research in fern biology. Future experiments that identify differentially expressed genes between wild-type and sex-determining mutants of Ceratopteris, such as *her1* and *her3* (Banks, 1994b, 1997d; Strain et al., 2001), should help refine the list of sex-determining genes. RNAi knock-down experiments are also underway to test the function of the genes hypothesized to be involved in sex determination.

LIST OF REFERENCES

LIST OF REFERENCES

Acosta, I. F., H. Laparra, S. P. Romero, E. Schmelz, M. Hamberg, J. P. Mottinger, M. A. Moreno and S. L. Dellaporta (2009). "tasselseed1 is a lipoxygenase affecting jasmonic acid signaling in sex determination of maize." Science **323**(5911): 262-265.

Allen, C. E. (1917). "A chromosome difference correlated with sex differences in *Sphaerocarpos*." Science **46**: 466-467.

Allen, C. E. (1919). "The Basis of Sex Inheritance in Spaerocarpos." Proc of the Am Philos Soc **58**(5): 289-316.

Alonso, J. M., T. Hirayama, G. Roman, S. Nourizadeh and J. R. Ecker (1999). "EIN2, a bifunctional transducer of ethylene and stress responses in Arabidopsis." Science **284**(5423): 2148-2152.

Anders, S. and W. Huber (2010). "Differential expression analysis for sequence count data." Genome Biol **11**(10): R106.

Aparicio, G., S. Gotz, A. Conesa, D. Segrelles, I. Blanquer, J. M. Garcia, V. Hernandez, M. Robles and M. Talon (2006). "Blast2GO goes grid: developing a grid-enabled prototype for functional genomics analysis." Stud Health Technol Inform **120**: 194-204.

Ashburner, M., C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin and G. Sherlock (2000). "Gene ontology: tool for the unification of biology. The Gene Ontology Consortium." Nat Genet **25**(1): 25-29.

Atallah, N. M. and J. A. Banks (2015). "Reproduction and the pheromonal regulation of sex type in fern gametophytes." Front Plant Sci **6**(100).
Atsmon, D. and E. Galun (1962). "Physiology of sex in Cucumis Sativus (L.) leaf age patterns and sexual differentiation of floral buds." Ann Bot **26**(2): 137-146.

Aya, K., Y. Hiwatashi, M. Kojima, H. Sakakibara, M. Ueguchi-Tanaka, M. Hasebe and M. Matsuoka (2011). "The Gibberellin perception system evolved to regulate a pre-existing GAMYB-mediated system during land plant evolution." Nat Commun **2**: 544.

Aya, K., M. Kobayashi, J. Tanaka, H. Ohyanagi, T. Suzuki, K. Yano, T. Takano, K. Yano and M. Matsuoka (2015). "De Novo Transcriptome Assembly of a Fern, Lygodium japonicum, and a Web Resource Database, Ljtrans DB." Plant Cell Physiol **56**(1): e5.

Aya, K., M. Ueguchi-Tanaka, M. Kondo, K. Hamada, K. Yano, M. Nishimura and M. Matsuoka (2009). "Gibberellin modulates anther development in rice via the transcriptional regulation of GAMYB." Plant Cell **21**(5): 1453-1472.

Bai, S. L., Y. B. Peng, J. X. Cui, H. T. Gu, L. Y. Xu, Y. Q. Li, Z. H. Xu and S. N. Bai (2004). "Developmental analyses reveal early arrests of the spore-bearing parts of reproductive organs in unisexual flowers of cucumber (Cucumis sativus L.)." Planta **220**(2): 230-240.

Bai, S. N. and Z. H. Xu (2012). "Bird-nest puzzle: can the study of unisexual flowers such as cucumber solve the problem of plant sex determination?" Protoplasma **249 Suppl 2**: S119-123.

Bai, S. N. and Z. H. Xu (2013). "Unisexual cucumber flowers, sex and sex differentiation." Int Rev Cell Mol Biol **304**: 1-55.

Ballottari, M., M. Mozzo, J. Girardon, R. Hienerwadel and R. Bassi (2013). "Chlorophyll triplet quenching and photoprotection in the higher plant monomeric antenna protein Lhcb5." J Phys Chem B **117**(38): 11337-11348.

Banks, J. A. (1993). "Mutations Affecting the Sexual Phenotype of the Ceratopteris-Richardii Gametophyte." J Cell Biochem: 13-13.

Banks, J. A. (1994). "Sex-Determining Genes in the Homosporous Fern Ceratopteris." Development **120**(7): 1949-1958.
Banks, J. A. (1997). "Sex determination in the fern Ceratopteris." Trends Plant Sci **2**(5): 175-180.

Banks, J. A. (1997). "The TRANSFORMER genes of the fern Ceratopteris simultaneously promote meristem and archegonia development and repress antheridia development in the developing gametophyte." Genetics **147**(4): 1885-1897.

Banks, J. A. (1999). "Gametophyte Development in Ferns." Annu Rev Plant Physiol Plant Mol Biol **50**: 163-186.

Banks, J. A., L. Hickok and M. A. Webb (1993). "The Programming of Sexual Phenotype in the Homosporous Fern Ceratopteris-Richardii." Int J Plant Sci **154**(4): 522-534.

Baroux, C., M. T. Raissig and U. Grossniklaus (2011). "Epigenetic regulation and reprogramming during gamete formation in plants." Curr Opin Genet Dev **21**(2): 124-133.

Bateman, R., DiMichele, W. (1994). "Heterospory: the most iterative key innovation in the evolutionary history of the plant kingdom." Biol Rev **69**: 345-417.

Benjamini, Y., D. Drai, G. Elmer, N. Kafkafi and I. Golani (2001). "Controlling the false discovery rate in behavior genetics research." Behav Brain Res **125**(1-2): 279-284.

Bensen, R. J., G. S. Johal, V. C. Crane, J. T. Tossberg, P. S. Schnable, R. B. Meeley and S. P. Briggs (1995). "Cloning and characterization of the maize An1 gene." Plant Cell **7**(1): 75-84.

Bergero, R. and D. Charlesworth (2011). "Preservation of the Y transcriptome in a 10-million-year-old plant sex chromosome system." Curr Biol **21**(17): 1470-1474.

Bergero, R., S. Qiu, A. Forrest, H. Borthwick and D. Charlesworth (2013). "Expansion of the pseudo-autosomal region and ongoing recombination suppression in the Silene latifolia sex chromosomes." Genetics **194**(3): 673-686.

Bernasconi, G., J. Antonovics, A. Biere, D. Charlesworth, L. F. Delph, D. Filatov, T. Giraud, M. E. Hood, G. A. Marais, D. McCauley, J. R. Pannell, J. A. Shykoff, B. Vyskot, L. M. Wolfe and A. Widmer (2009). "Silene as a model system in ecology and evolution." Heredity (Edinb) **103**(1): 5-14.

Berr, A., E. J. McCallum, R. Menard, D. Meyer, J. Fuchs, A. Dong and W. H. Shen (2010). "Arabidopsis SET DOMAIN GROUP2 is required for H3K4 trimethylation and is crucial for both sporophyte and gametophyte development." Plant Cell **22**(10): 3232-3248.

Blackburn, K. B. (1923). "Sex chromosomes in plants." Nature **112**: 687-688.

Bonnet, O. T. (1940). "Development of the staminate and pistillate inflorescences of sweet corn." J Agric Res **60**: 25-37.

Borges, F., J. P. Calarco and R. A. Martienssen (2012). "Reprogramming the epigenome in Arabidopsis pollen." CSH Symp Quant Biol **77**: 1-5.

Boualem, A., C. Troadec, I. Kovalski, M. A. Sari, R. Perl-Treves and A. Bendahmane (2009). "A conserved ethylene biosynthesis enzyme leads to andromonoecy in two cucumis species." PLoS One **4**(7): e6144.

Calarco, J. P., F. Borges, M. T. Donoghue, F. Van Ex, P. E. Jullien, T. Lopes, R. Gardner, F. Berger, J. A. Feijo, J. D. Becker and R. A. Martienssen (2012). "Reprogramming of DNA methylation in pollen guides epigenetic inheritance via small RNA." Cell **151**(1): 194-205.

Calderon-Urrea, A. and S. L. Dellaporta (1999). "Cell death and cell protection genes determine the fate of pistils in maize." Development **126**(3): 435-441.

Charlesworth, D. (2002). "Plant sex determination and sex chromosomes." Heredity **88**(2): 94-101.

Tarailo-Graovac, M. and N. Chen (2004). "Using RepeatMasker to identify repetitive elements in genomic sequences." Curr Protoc Bioinformatics. Chapter 4: Unit 4.10.

Cheng, P. C., R. I. Greyson and D. B. Walden (1983). "Organ Initiation and the Development of Unisexual Flowers in the Tassel and Ear of Zea mays." Am J Bot**70**(3): 450-462.

Chettoor, A. M., S. A. Givan, R. A. Cole, C. T. Coker, E. Unger-Wallace, Z. Vejlupkova, E. Vollbrecht, J. E. Fowler and M. M. Evans (2014). "Discovery of novel transcripts and gametophytic functions via RNA-seq analysis of maize gametophytic transcriptomes." Genome Biol **15**(7): 414.

Chiu, J. C., E. K. Lee, M. G. Egan, I. N. Sarkar, G. M. Coruzzi and R. DeSalle (2006). "OrthologID: automation of genome-scale ortholog identification within a parsimony framework." Bioinformatics **22**(6): 699-707.

Chuck, G., R. Meeley, E. Irish, H. Sakai and S. Hake (2007). "The maize tasselseed4 microRNA controls sex determination and meristem cell fate by targeting Tasselseed6/indeterminate spikelet1." Nat Genet **39**(12): 1517-1521.

Chun, P. T. and L. G. Hickok (1992). "Inheritance of 2 Mutations Conferring Glyphosate Tolerance in the Fern Ceratopteris-Richardii." Can J Bot **70**(5): 1097-1099.

Clapier, C. R. and B. R. Cairns (2009). "The biology of chromatin remodeling complexes." Annu Rev Biochem **78**: 273-304.

Conesa, A. and S. Gotz (2008). "Blast2GO: A comprehensive suite for functional analysis in plant genomics." Int J Plant Genomics **2008**: 619832.

Conesa, A., S. Gotz, J. M. Garcia-Gomez, J. Terol, M. Talon and M. Robles (2005). "Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research." Bioinformatics **21**(18): 3674-3676.

Cordle, A. R., E. E. Irish and C. L. Cheng (2007). "Apogamy Induction in Ceratopteris richardii." Int. J. Plant Sci. **168**: 361-369.

Cordle, A. R., E. E. Irish and C. L. Cheng (2012). "Gene expression associated with apogamy commitment in Ceratopteris richardii." Sex Plant Reprod **25**(4): 293-304.

Cutler, S. R., P. L. Rodriguez, R. R. Finkelstein and S. R. Abrams (2010). "Abscisic acid: emergence of a core signaling network." Annu Rev Plant Biol **61**: 651-679.

Daviere, J. M. and P. Achard (2013). "Gibberellin signaling in plants." Development **140**(6): 1147-1151.

Davidson, N. and Oshlack, A (2014). "Corset: enabling differential gene expression analysis for *de novo* assembled transcriptomes." Genome Biol **15**: 410.

DeLong, A., A. Calderon-Urrea and S. L. Dellaporta (1993). "Sex determination gene TASSELSEED2 of maize encodes a short-chain alcohol dehydrogenase required for stage-specific floral organ abortion." Cell **74**(4): 757-768.

Der, J. P., M. S. Barker, N. J. Wickett, C. W. dePamphilis and P. G. Wolf (2011). "De novo characterization of the gametophyte transcriptome in bracken fern, Pteridium aquilinum." BMC Genomics **12**: 99.

Desfeux, C., S. Maurice, J. P. Henry, B. Lejeune and P. H. Gouyon (1996). "Evolution of reproductive systems in the genus Silene." Proc Biol Sci **263**(1369): 409-414.

DeYoung, B., T. Weber, B. Hass and J. A. Banks (1997). "Generating autotetraploid sporophytes and their use in analyzing mutations affecting gametophyte development in the fern Ceratopteris." Genetics **147**(2): 809-814.

Döpp, W. (1950). "Eine die Antheridienbildung bei Farnen fördernde Substanz in den Prothallien von *Pteridium aquilinum* (L.)." Kuhn Ber. Deut. Botan. Ges. **63**: 139-146.

Döpp, W. (1950). "Eine die Antheridienbildung bei Farnen fördernde Substanz in den Prothallien von Pteridium aquilinum L. Kun. Ber. Dtsch." Bot. Gaz. **63**: 139-147.

Earl, D., K. Bradnam, J. St John, A. Darling, D. Lin, J. Fass, H. O. Yu, V. Buffalo, D. R. Zerbino, M. Diekhans, N. Nguyen, P. N. Ariyaratne, W. K. Sung, Z. Ning, M. Haimel, J. T. Simpson, N. A. Fonseca, I. Birol, T. R. Docking, I. Y. Ho, D. S. Rokhsar, R. Chikhi, D. Lavenier, G. Chapuis, D. Naquin, N. Maillet, M. C. Schatz, D. R. Kelley, A. M. Phillippy, S. Koren, S. P. Yang, W. Wu, W. C. Chou, A. Srivastava, T. I. Shaw, J. G. Ruby, P. Skewes-Cox, M. Betegon, M. T. Dimon, V. Solovyev, I. Seledtsov, P. Kosarev, D. Vorobyev, R. Ramirez-Gonzalez, R. Leggett, D. MacLean, F. Xia, R. Luo, Z. Li, Y. Xie, B. Liu, S. Gnerre, I. MacCallum, D. Przybylski, F. J. Ribeiro, S. Yin, T. Sharpe, G. Hall, P. J. Kersey, R. Durbin, S. D. Jackman, J. A. Chapman, X. Huang, J. L. DeRisi, M. Caccamo, Y. Li, D. B. Jaffe, R. E. Green, D. Haussler, I. Korf and B. Paten (2011). "Assemblathon 1: a competitive assessment of de novo short read assembly methods." Genome Res **21**(12): 2224-2241.

Ebbs, M. L. and J. Bender (2006). "Locus-specific control of DNA methylation by the Arabidopsis SUVH5 histone methyltransferase." Plant Cell **18**(5): 1166-1176.

Eberle, J. R. and J. A. Banks (1996). "Genetic interactions among sex-determining genes in the fern Ceratopteris richardii." Genetics **142**(3): 973-985.

Egan, M., E. K. Lee, J. C. Chiu, G. Coruzzi and R. Desalle (2009). "Gene orthology assessment with OrthologID." Methods Mol Biol **537**: 23-38.

Emerson, R. A., G. W. Beadle and A. C. Fraser (1935). "A Summary of Linkage Studies in Maize." Cornell University Agricultural Experiment Station **180**: 1-83.

Farbos, I., M. Oliveira, I. Negrutiu and A. Mouras (1997). "Sex organ determination and differentiation in the dioecious plant *Melandrium album* (*Silene latifolia*): a cytological and histological analysis." Sexual Plant Reproduction **10**: 155-167.

Farbos, I., J. Veuskens, B. Vyskot, M. Oliveira, S. Hinnisdaels, A. Aghmir, A. Mouras and I. Negrutiu (1999). "Sexual dimorphism in white campion: deletion on the Y chromosome results in a floral asexual phenotype." Genetics **151**(3): 1187-1196.

Farrar, D. R. and J. T. Mickel (1991). "SocietyVittaria appalachiana: A Name for the "Appalachian Gametophyte"." Am Fern J **81**: 69-75.

Farrona, S., L. Hurtado, J. L. Bowman and J. C. Reyes (2004). "The Arabidopsis thaliana SNF2 homolog AtBRM controls shoot development and flowering." Development **131**(20): 4965-4975.

Fernando, D. D., J. N. Owens, X. S. Yu and A. K. M. Ekramoddoullah (2001). "RNA and protein synthesis during in vitro pollen germination and tube elongation in Pinus monticola and other conifers." Sexual Plant Reproduction **13**(5): 259-264.

Filatov, D. A. (2005). "Evolutionary history of Silene latifolia sex chromosomes revealed by genetic mapping of four genes." Genetics **170**(2): 975-979.

Filatov, D. A. (2005). "Isolation of genes from plant Y chromosomes." Methods Enzymol **395**: 418-442.

Finn, R. D., J. Clements and S. R. Eddy (2011). "HMMER web server: interactive sequence similarity searching." Nucleic Acids Res **39**(Web Server issue): W29-37.

Fleet, C. M. and T. P. Sun (2005). "A DELLAcate balance: the role of gibberellin in plant morphogenesis." Curr Opin Plant Biol **8**(1): 77-85.

Fujioka, S., H. Yamane, C. R. Spray, P. Gaskin, J. Macmillan, B. O. Phinney and N. Takahashi (1988). "Qualitative and Quantitative Analyses of Gibberellins in Vegetative Shoots of Normal, dwarf-1, dwarf-2, dwarf-3, and dwarf-5 Seedlings of Zea mays L." Plant Physiol **88**(4): 1367-1372.

Fujita, Y., M. Fujita, K. Shinozaki and K. Yamaguchi-Shinozaki (2011). "ABA-mediated transcriptional regulation in response to osmotic stress in plants." J Plant Res **124**(4): 509-525.

Furber, M., L. Mander, J. Nester, N. Takahashi and H. Yamane (1989). "Structure of a novel antheridiogen from the fern Anemia mexicana." Phytochem. **28**(63-66).

Galun, E. (1961). "Study of the inheritance of sex expression in the cucumber: the interaction of major genes with modifying genetic and non-genetic factors." Genetica **32**: 134-163.

Gil, P., E. Dewey, J. Friml, Y. Zhao, K. C. Snowden, J. Putterill, K. Palme, M. Estelle and J. Chory (2001). "BIG: a calossin-like protein required for polar auxin transport in Arabidopsis." Genes Dev **15**(15): 1985-1997.

Gocal, G. F., A. T. Poole, F. Gubler, R. J. Watts, C. Blundell and R. W. King (1999). "Long-day up-regulation of a GAMYB gene during Lolium temulentum inflorescence formation." Plant Physiol **119**(4): 1271-1278.

Gocal, G. F., C. C. Sheldon, F. Gubler, T. Moritz, D. J. Bagnall, C. P. MacMillan, S. F. Li, R. W. Parish, E. S. Dennis, D. Weigel and R. W. King (2001). "GAMYB-like genes, flowering, and gibberellin signaling in Arabidopsis." Plant Physiol **127**(4): 1682-1693.

Goff, S. A., M. Vaughn, S. McKay, E. Lyons, A. E. Stapleton, D. Gessler, N. Matasci, L. Wang, M. Hanlon, A. Lenards, A. Muir, N. Merchant, S. Lowry, S. Mock, M. Helmke, A. Kubach, M. Narro, N. Hopkins, D. Micklos, U. Hilgert, M. Gonzales, C. Jordan, E. Skidmore, R. Dooley, J. Cazes, R. McLay, Z. Lu, S. Pasternak, L. Koesterke, W. H. Piel, R. Grene, C. Noutsos, K. Gendler, X. Feng, C. Tang, M. Lent, S. J. Kim, K. Kvilekval, B. S. Manjunath, V. Tannen, A. Stamatakis, M. Sanderson, S. M. Welch, K. A. Cranston, P. Soltis, D. Soltis, B. O'Meara, C. Ane, T. Brutnell, D. J. Kleibenstein, J. W. White, J. Leebens-Mack, M. J. Donoghue, E. P. Spalding, T. J. Vision, C. R. Myers, D. Lowenthal, B. J. Enquist, B. Boyle, A. Akoglu, G. Andrews, S. Ram, D. Ware, L. Stein and D. Stanzione (2011). "The iPlant Collaborative: Cyberinfrastructure for Plant Biology." Front Plant Sci **2**: 34.

Gong, Z., T. Morales-Ruiz, R. R. Ariza, T. Roldan-Arjona, L. David and J. K. Zhu (2002). "ROS1, a repressor of transcriptional gene silencing in Arabidopsis, encodes a DNA glycosylase/lyase." Cell **111**(6): 803-814.

Goodstein, D. M., S. Shu, R. Howson, R. Neupane, R. D. Hayes, J. Fazo, T. Mitros, W. Dirks, U. Hellsten, N. Putnam and D. S. Rokhsar (2012). "Phytozome: a comparative platform for green plant genomics." Nucleic Acids Res **40**(Database issue): D1178-1186.

Gotz, S., J. M. Garcia-Gomez, J. Terol, T. D. Williams, S. H. Nagaraj, M. J. Nueda, M. Robles, M. Talon, J. Dopazo and A. Conesa (2008). "High-throughput functional annotation and data mining with the Blast2GO suite." Nucleic Acids Res **36**(10): 3420-3435.

Gou, J., S. H. Strauss, C. J. Tsai, K. Fang, Y. Chen, X. Jiang and V. B. Busov (2010). "Gibberellins regulate lateral root formation in Populus through interactions with auxin and other hormones." Plant Cell **22**(3): 623-639.

Grabherr, M. G., B. J. Haas, M. Yassour, J. Z. Levin, D. A. Thompson, I. Amit, X. Adiconis, L. Fan, R. Raychowdhury, Q. Zeng, Z. Chen, E. Mauceli, N. Hacohen, A. Gnirke, N. Rhind, F. di Palma, B. W. Birren, C. Nusbaum, K. Lindblad-Toh, N. Friedman and A. Regev (2011). "Full-length transcriptome assembly from RNA-Seq data without a reference genome." Nat Biotechnol **29**(7): 644-652.

Grant, S., B. Hunkirchen and S. Heinz (1994). "Developmental differences between male and female flowers in the dioecious plant *Silene latifolia*." The Plant Journal **6**(4): 471-480.

Grini, P. E., T. Thorstensen, V. Alm, G. Vizcay-Barrena, S. S. Windju, T. S. Jorstad, Z. A. Wilson and R. B. Aalen (2009). "The ASH1 HOMOLOG 2 (ASHH2) histone H3 methyltransferase is required for ovule and anther development in Arabidopsis." PLoS One **4**(11): e7817.

Gubler, F., P. M. Chandler, R. G. White, D. J. Llewellyn and J. V. Jacobsen (2002). "Gibberellin signaling in barley aleurone cells. Control of SLN1 and GAMYB expression." Plant Physiol **129**(1): 191-200.

Gubler, F., R. Kalla, J. K. Roberts and J. V. Jacobsen (1995). "Gibberellin-regulated expression of a myb gene in barley aleurone cells: evidence for Myb transactivation of a high-pI alpha-amylase gene promoter." Plant Cell **7**(11): 1879-1891.

Gubler, F., D. Raventos, M. Keys, R. Watts, J. Mundy and J. V. Jacobsen (1999). "Target genes and regulatory domains of the GAMYB transcriptional activator in cereal aleurone." Plant J **17**(1): 1-9.

Hao, Y. J., D. H. Wang, Y. B. Peng, S. L. Bai, L. Y. Xu, Y. Q. Li, Z. H. Xu and S. N. Bai (2003). "DNA damage in the early primordial anther is closely correlated with stamen arrest in the female flower of cucumber (Cucumis sativus L.)." Planta **217**(6): 888-895.

Hartwig, T., G. S. Chuck, S. Fujioka, A. Klempien, R. Weizbauer, D. P. Potluri, S. Choe, G. S. Johal and B. Schulz (2011). "Brassinosteroid control of sex determination in maize." Proc Natl Acad Sci U S A **108**(49): 19814-19819.

Hedden, P. and S. G. Thomas (2012). "Gibberellin biosynthesis and its regulation." Biochem J **444**(1): 11-25.

Henderson, I. R., X. Zhang, C. Lu, L. Johnson, B. C. Meyers, P. J. Green and S. E. Jacobsen (2006). "Dissecting Arabidopsis thaliana DICER function in small RNA processing, gene silencing and DNA methylation patterning." Nat Genet **38**(6): 721-725.

Henderson, J. T., H. C. Li, S. D. Rider, A. P. Mordhorst, J. Romero-Severson, J. C. Cheng, J. Robey, Z. R. Sung, S. C. de Vries and J. Ogas (2004). "PICKLE acts throughout the plant to repress expression of embryonic traits and may play a role in gibberellin-dependent responses." Plant Physiol **134**(3): 995-1005.

Hickok, L. G. (1977). "Apomictic Mutant for Sticky Chromosomes in Fern Ceratopteris." Can. J. Bot **55**(16): 2186-2195.

Hickok, L. G. (1983). "Abscisic acid blocks antheridiogen-induced antheridium formation in gametophytes of the fern Ceratopteris." Can. J. Bot. **63**: 888-892.

Hickok, L. G. (1985). "Abscisic-Acid Resistant Mutants in the Fern Ceratopteris - Characterization and Genetic-Analysis." Can. J. Bot. **63**(9): 1582-1585.

Hickok, L. G. and O. J. Schwarz (1989). "Genetic-Characterization of a Mutation That Enhances Paraquat Tolerance in the Fern Ceratopteris-Richardii." Theor Appl Genet **77**(2): 200-204.

Hickok, L. G., R. J. Scott and T. R. Warne (1985). "Isolation and Characterization of Antheridiogen-Resistant Mutants in the Fern Ceratopteris." Am J Bot **72**(6): 922-922.

Hickok, L. G., D. L. Vogelien and T. R. Warne (1991). "Selection of a Mutation Conferring High Nacl Tolerance to Gametophytes of Ceratopteris." Theor Appl Genet **81**(3): 293-300.

Hickok, L. G., T. R. Warne and R. S. Fribourg (1995). "The Biology of the Fern Ceratopteris and Its Use as a Model System." Int J Plant Sci **156**(3): 332-345.

Hickok, L. G., T. R. Warne and M. K. Slocum (1987). "Ceratopteris richardii - Applications for Experimental Plant Biology." Am. J. Bot. **74**: 1304-1316.

Hickok, L. G., T. R. Warne and M. K. Slocum (1987). "Ceratopteris-Richardii - Applications for Experimental Plant Biology." Am. J. Bot. **74**(8): 1304-1316.

Hirano, K., M. Nakajima, K. Asano, T. Nishiyama, H. Sakakibara, M. Kojima, E. Katoh, H. Xiang, T. Tanahashi, M. Hasebe, J. A. Banks, M. Ashikari, H. Kitano, M. Ueguchi-Tanaka and M. Matsuoka (2007). "The GID1-mediated gibberellin perception mechanism is conserved in the Lycophyte Selaginella moellendorffii but not in the Bryophyte Physcomitrella patens." Plant Cell **19**(10): 3058-3079.

Hochberg, Y. and Y. Benjamini (1990). "More powerful procedures for multiple significance testing." Stat Med **9**(7): 811-818.

Huanca-Mamani, W., M. Garcia-Aguilar, G. Leon-Martinez, U. Grossniklaus and J. P. Vielle-Calzada (2005). "CHR11, a chromatin-remodeling factor essential for nuclear proliferation during female gametogenesis in Arabidopsis thaliana." PNAS **102**(47): 17231-17236.

Huson, D. H., S. Mitra, H. J. Ruscheweyh, N. Weber and S. C. Schuster (2011). "Integrative analysis of environmental sequences using MEGAN4." Genome Res **21**(9): 1552-1560.

Hwang, I., J. Sheen and B. Muller (2012). "Cytokinin signaling networks." Annu Rev Plant Biol **63**: 353-380.

Ihnatowicz, A., P. Pesaresi and D. Leister (2007). "The E subunit of photosystem I is not essential for linear electron flow and photoautotrophic growth in Arabidopsis thaliana." Planta **226**(4): 889-895.

Irish, E. E. (1999). Maize sex determination Sex Determination in Plants. C. Ainsworth. Oxford, UK, BIOS Scientific Publishers**:** 183-188.

Jackson, J. P., A. M. Lindroth, X. Cao and S. E. Jacobsen (2002). "Control of CpNpG DNA methylation by the KRYPTONITE histone H3 methyltransferase." Nature **416**(6880): 556-560.

Jiang, F., M. Yang, W. Guo, X. Wang and L. Kang (2012). "Large-scale transcriptome analysis of retroelements in the migratory locust, Locusta migratoria." PLoS One **7**(7): e40532.

Jiang, L., A. J. Wijeratne, S. Wijeratne, M. Fraga, T. Meulia, D. Doohan, Z. Li and F. Qu (2013). "Profiling mRNAs of two Cuscuta species reveals possible candidate transcripts shared by parasitic plants." PLoS One **8**(11): e81389.

Jimenez, A., L. G. Quintanilla, S. Pajaron and E. Pangua (2008). "Reproductive and competitive interactions among gametophytes of the allotetraploid fern Dryopteris corleyi and its two diploid parents." Annals of Botany **102**(3): 353-359.

Jones, D. F. (1925). "Heritable characters in maize. XXIII. Silkless." Journal of Heredity **16**: 339-341.

Jones, P., D. Binns, H. Y. Chang, M. Fraser, W. Li, C. McAnulla, H. McWilliam, J. Maslen, A. Mitchell, G. Nuka, S. Pesseat, A. F. Quinn, A. Sangrador-Vegas, M. Scheremetjew, S. Y. Yong, R. Lopez and S. Hunter (2014). "InterProScan 5: genome-scale protein function classification." Bioinformatics **30**(9): 1236-1240.

Jullien, P. E., D. Susaki, R. Yelagandula, T. Higashiyama and F. Berger (2012). "DNA methylation dynamics during sexual reproduction in Arabidopsis thaliana." Curr Biol **22**(19): 1825-1830.

Kaiser, V. B., R. Bergero and D. Charlesworth (2009). "Slcyt, a newly identified sex-linked gene, has recently moved onto the X chromosome in Silene latifolia (Caryophyllaceae)." Mol Biol Evol **26**(10): 2343-2351.

Kanehisa, M., S. Goto, Y. Sato, M. Furumichi and M. Tanabe (2012). "KEGG for integration and interpretation of large-scale molecular data sets." Nucleic Acids Res **40**(Database issue): D109-114.

Kaneko, M., Y. Inukai, M. Ueguchi-Tanaka, H. Itoh, T. Izawa, Y. Kobayashi, T. Hattori, A. Miyao, H. Hirochika, M. Ashikari and M. Matsuoka (2004). "Loss-of-function mutations of the rice GAMYB gene impair alpha-amylase expression in aleurone and flower development." Plant Cell **16**(1): 33-44.

Kater, M. M., J. Franken, K. J. Carney, L. Colombo and G. C. Angenent (2001). "Sex determination in the monoecious species cucumber is confined to specific floral whorls." Plant Cell **13**(3): 481-493.

Kellogg, E. A. and J. A. Birchler (1993). "Linking Phylogeny and Genetics: *Zea Mays* as a Tool for Phylogenetic Studies." Systematic Biology **42**(4): 415-439.

Kim, J. C., H. Laparra, A. Calderon-Urrea, J. P. Mottinger, M. A. Moreno and S. L. Dellaporta (2007). "Cell cycle arrest of stamen initials in maize sex determination." Genetics **177**(4): 2547-2551.

Knopf, R. R. and T. Trebitsh (2006). "The female-specific Cs-ACS1G gene of cucumber. A case of gene duplication and recombination between the non-sex-specific 1-aminocyclopropane-1-carboxylate synthase gene and a branched-chain amino acid transaminase gene." Plant Cell Physiol **47**(9): 1217-1228.

Koizumi, A., K. Yamanaka, K. Nishihara, Y. Kazama, T. Abe and S. Kawano (2010). "Two separate pathways including SlCLV1, SlSTM and SlCUC that control carpel development in a bisexual mutant of Silene latifolia." Plant Cell Physiol **51**(2): 282-293.

Kong, L., Y. Zhang, Z. Q. Ye, X. Q. Liu, S. Q. Zhao, L. Wei and G. Gao (2007). "CPC: assess the protein-coding potential of transcripts using sequence features and support vector machine." Nucleic Acids Res **35**(Web Server issue): W345-349.

Kozik, A., M. Matvienko, I. Kozik, H. Van Leeuwen, A. Van Deynze and R. Michelmore (2008). Eukaryotic ultra conserved orthologs and estimation of gene capture In EST libraries. Plant and Animal Genomes Conference **16**.

Krogh, A., B. Larsson, G. von Heijne and E. L. Sonnhammer (2001). "Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes." J Mol Biol **305**(3): 567-580.

Kubicki, B. (1969). "Comparative studies on sex determination in cucumber (*Cucumis sativus* L.) and muskmelon (*Cucumis melo* L.)." Genetica Polonica **10**: 167-183.

Kubicki, B. (1969). "Investigations on sex determination in cucumber (*Cucumis sativus* L.). V. Genes controlling intensity of femaleness." Genetica Polonica **10**: 69-85.

Kubicki, B. (1969). "Investigations on sex determination in cucumber (*Cucumis sativus* L.). VII. Andromonoecious and hermaphroditism." Genetica Polonica **10**: 101-120.

Kurumatani, M., K. Yagi, T. Murata, M. Tezuka, L. N. Mander, M. Nishiyama and H. Yama (2001). "Isolation and identification of antheridiogens in the ferns, Lygodium microphyllum and Lygodium reticulatum." Biosci Biotechnol Biochem **65**(10): 2311-2314.

Kushiro, T., M. Okamoto, K. Nakabayashi, K. Yamagishi, S. Kitamura, T. Asami, N. Hirai, T. Koshiba, Y. Kamiya and E. Nambara (2004). "The Arabidopsis cytochrome P450 CYP707A encodes ABA 8'-hydroxylases: key enzymes in ABA catabolism." EMBO J **23**(7): 1647-1656.

Lardon, A., S. Georgiev, A. Aghmir, G. Le Merrer and I. Negrutiu (1999). "Sexual Dimorphism in White Campion: Complex Control of Carpel Number Is Revealed by Y Chromosome Deletions." Genetics **151**(3): 1173-1185.

Law, J. A. and S. E. Jacobsen (2010). "Establishing, maintaining and modifying DNA methylation patterns in plants and animals." Nat Rev Genet **11**(3): 204-220.

Lebel-Hardenack, S., E. Hauser, T. F. Law, J. Schmid and S. R. Grant (2002). "Mapping of sex determination loci on the white campion (Silene latifolia) Y chromosome using amplified fragment length polymorphism." Genetics **160**(2): 717-725.

Leng, N., J. A. Dawson, J. A. Thomson, V. Ruotti, A. I. Rissman, B. M. Smits, J. D. Haag, M. N. Gould, R. M. Stewart and C. Kendziorski (2013). "EBSeq: an empirical Bayes hierarchical model for inference in RNA-seq experiments." Bioinformatics **29**(8): 1035-1043.

Li, B. and C. N. Dewey (2011). "RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome." BMC bioinformatics **12**: 323.

Li, B., V. Ruotti, R. M. Stewart, J. A. Thomson and C. N. Dewey (2010). "RNA-Seq gene expression estimation with read mapping uncertainty." Bioinformatics **26**(4): 493-500.

Li, Z., S. Huang, S. Liu, J. Pan, Z. Zhang, Q. Tao, Q. Shi, Z. Jia, W. Zhang, H. Chen, L. Si, L. Zhu and R. Cai (2009). "Molecular isolation of the M gene suggests that a conserved-residue conversion induces the formation of bisexual flowers in cucumber plants." Genetics **182**(4): 1381-1385.

Lisch, D. and J. L. Bennetzen (2011). "Transposable element origins of epigenetic gene regulation." Curr Opin Plant Biol **14**(2): 156-161.

Livak, K. J. and T. D. Schmittgen (2001). "Analysis of relative gene expression data using real-time quantitative PCR and the 2(-Delta Delta C(T)) Method." Methods **25**(4): 402-408.

Livingston, A. K., J. A. Cruz, K. Kohzuma, A. Dhingra and D. M. Kramer (2010). "An Arabidopsis mutant with high cyclic electron flow around photosystem I (hcef) involving the NADPH dehydrogenase complex." Plant Cell **22**(1): 221-233.

Lohse, M., A. M. Bolger, A. Nagel, A. R. Fernie, J. E. Lunn, M. Stitt and B. Usadel (2012). "RobiNA: a user-friendly, integrated software solution for RNA-Seq-based transcriptomics." Nucleic Acids Res **40**(Web Server issue): W622-627.

Lohse, M., A. M. Bolger, A. Nagel, A. R. Fernie, J. E. Lunn, M. Stitt and B. Usadel (2012). "RobiNA: a user-friendly, integrated software solution for RNA-Seq-based transcriptomics." Nucleic Acids Res **40**(W1): W622-W627.

Loraine, A. E., S. McCormick, A. Estrada, K. Patel and P. Qin (2013). "RNA-seq of Arabidopsis pollen uncovers novel transcription and alternative splicing." Plant Physiol **162**(2): 1092-1109.

Lorbeer, G. (1934). "Die Zytologie der Lebermoose mit besonderer Berucksichtingung allgemeiner Chromosomenfragen." Jahrb. Wiss. Bot. **80**: 567-817.

Love, M. I., W. Huber and S. Anders (2014). "Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2." Genome Biol **15**(12): 550.

Malepszy, S. and K. Niemirowicz-Szczytt (1991). "Sex determination in cucumber (*Cucumis sativus*) as a model system for molecular biology." Plant Science **80**: 39-47.

Martienssen, R. and V. Chandler (2013). Molecular mechanisms of transposon epigenetic regulation. Plant Transposons and Genome Dynamics in Evolution. N. Fedoroff. Ames, Iowa, Wiley-Blackwell**:** 71-92.

Matzke, M. A. and R. A. Mosher (2014). "RNA-directed DNA methylation: an epigenetic pathway of increasing complexity." Nat Rev Genet **15**(6): 394-408.

Mibus, H. and T. Tatlioglu (2004). "Molecular characterization and isolation of the F/f gene for femaleness in cucumber (Cucumis sativus L.)." Theor Appl Genet **109**(8): 1669-1676.

Moneger, F., N. Barbacar and I. Negrutiu (2000). "Dioecious *Silene* at the X-road: the reasons." Sexual Plant Reproduction **12**: 245-249.

Munoz-Bertomeu, J., B. Cascales-Minana, J. M. Mulet, E. Baroja-Fernandez, J. Pozueta-Romero, J. M. Kuhn, J. Segura and R. Ros (2009). "Plastidial glyceraldehyde-3-phosphate dehydrogenase deficiency leads to altered root development and affects the sugar and amino acid balance in Arabidopsis." Plant Physiol **151**(2): 541-558.

Naf, U. (1979). Antheridiogens and antheridial development. The Experimental Biology of Ferns. A. F. Dyer. New York, Academic Press**:** 436-470.

Näf, U. (1959). "Control of Antheridium Formation in the Fern Species *Anemia Phyllitides*." Nature **184**(4689): 798-800.

Nakashima, K., Y. Fujita, K. Katsura, K. Maruyama, Y. Narusaka, M. Seki, K. Shinozaki and K. Yamaguchi-Shinozaki (2006). "Transcriptional regulation of ABI3- and ABA-responsive genes including RD29B and RD29A in seeds, germinating embryos, and seedlings of Arabidopsis." Plant Mol Biol **60**(1): 51-68.

Newhouse, S. and D. To (2010). cmpfastq. http://compbio.brc.iop.kcl.ac.uk/software/download/cmpfastq.

Neyman, J. and Pearson, E.S. (1928) On the use and interpretation of certain test criteria for purposes of statistical inference. *Biometrika*, 20, 175-240, 263-294.

Nickerson, N. H. and E. E. Dale (1955). "Tassel Modifications in Zea Mays." Annals of the Missouri Botanical Garden **42**(3): 195-211.

Noh, B., A. S. Murphy and E. P. Spalding (2001). "Multidrug resistance-like genes of Arabidopsis required for auxin transport and auxin-mediated development." Plant Cell **13**(11): 2441-2454.

Ogas, J., S. Kaufmann, J. Henderson and C. Somerville (1999). "PICKLE is a CHD3 chromatin-remodeling factor that regulates the transition from embryonic to vegetative development in Arabidopsis." PNAS **96**(24): 13839-13844.

Okano, Y., N. Aono, Y. Hiwatashi, T. Murata, T. Nishiyama, T. Ishikawa, M. Kubo and M. Hasebe (2009). "A polycomb repressive complex 2 gene regulates apogamy and gives evolutionary insights into early land plant evolution." PNAS **106**(38): 16321-16326.

Onodera, Y., J. R. Haag, T. Ream, P. Costa Nunes, O. Pontes and C. S. Pikaard (2005). "Plant nuclear RNA polymerase IV mediates siRNA and DNA methylation-dependent heterochromatin formation." Cell **120**(5): 613-622.

Peng, J., D. E. Richards, N. M. Hartley, G. P. Murphy, K. M. Devos, J. E. Flintham, J. Beales, L. J. Fish, A. J. Worland, F. Pelica, D. Sudhakar, P. Christou, J. W. Snape, M. D. Gale and N. P. Harberd (1999). "'Green revolution' genes encode mutant gibberellin response modulators." Nature **400**(6741): 256-261.

Peng, M., Y. Cui, Y. M. Bi and S. J. Rothstein (2006). "AtMBD9: a protein with a methyl-CpG-binding domain regulates flowering time and shoot branching in Arabidopsis." Plant J **46**(2): 282-296.

Perl-Treves, R. (1999). Male to female conversion along the cucumber shoot: Approaches to studying sex genes and floral development in *Cucumis sativus*. Sex Determination in Plants. C. Ainsworth. Oxford, UK, Bios Scientific Publishers**:** 189-216.

Perl-Treves, R. and P. A. Rajagopalan (2006). Close, yet separate: patterns of male and female floral development in monecious species. Flower Development and Manipulation. C. Ainsworth. Blackwell, Oxford**:** 117-146.

Petersen, T. N., S. Brunak, G. von Heijne and H. Nielsen (2011). "SignalP 4.0: discriminating signal peptides from transmembrane regions." Nat Methods **8**(10): 785-786.

Petrasek, J., J. Mravec, R. Bouchard, J. J. Blakeslee, M. Abas, D. Seifertova, J. Wisniewska, Z. Tadele, M. Kubes, M. Covanova, P. Dhonukshe, P. Skupa, E. Benkova, L. Perry, P. Krecek, O. R. Lee, G. R. Fink, M. Geisler, A. S. Murphy, C. Luschnig, E. Zazimalova and J. Friml (2006). "PIN proteins perform a rate-limiting function in cellular auxin efflux." Science **312**(5775): 914-918.

Phinney, B. O. (1982). Chemical genetics and the gibberellin pathway in *Zea mays* L. Plant growth substances 1982. P. F. Wareing. London ; New York, Academic Press**:** 101-110.

Phipps, I. F. (1928). "Heritable characters in maize. XXXI Tassel-seed 4." Journal of Heredity **19**: 399-404.

Plackett, A. R., L. Huang, H. L. Sanders and J. A. Langdale (2014). "High-efficiency stable transformation of the model fern species Ceratopteris richardii via microparticle bombardment." Plant Physiol **165**(1): 3-14.

Pontier, D., C. Picart, F. Roudier, D. Garcia, S. Lahmy, J. Azevedo, E. Alart, M. Laudie, W. M. Karlowski, R. Cooke, V. Colot, O. Voinnet and T. Lagrange (2012). "NERD, a plant-specific GW protein, defines an additional RNAi-dependent chromatin-based pathway in Arabidopsis." Mol Cell **48**(1): 121-132.

Prantl, K. A. E. (1881). "Beobachtungen üer die Ernährung der Farnprothallien und die Verteilung der Sexual Organe: Apogamie." Just's Bot. Jahrb. **15**: 553-574.

Qian, W., D. Miki, H. Zhang, Y. Liu, X. Zhang, K. Tang, Y. Kan, H. La, X. Li, S. Li, X. Zhu, X. Shi, K. Zhang, O. Pontes, X. Chen, R. Liu, Z. Gong and J. K. Zhu (2012). "A histone acetyltransferase regulates active DNA demethylation in Arabidopsis." Science **336**(6087): 1445-1448.

Quevillon, E., V. Silventoinen, S. Pillai, N. Harte, N. Mulder, R. Apweiler and R. Lopez (2005). "InterProScan: protein domains identifier." Nucleic Acids Res **33**(Web Server issue): W116-120.

Raghavan, V. (1965). "Actinomycin D: its effects on two-dimensional growth in fern gametophytes." Exp Cell Res **39**(2): 689-692.

Raghavan, V. (1968). "Actinomycin D-Induced Changes in Growth and Ribonucleic Acid Metabolism in the Gametophytes of Bracken Fern." Am. J. Bot. **55**(7): 767-772.

Raghavan, V. (1970). "Germination of bracken fern spores. Regulation of protein and RNA synthesis during initiation and growth of the rhizoid." Exp Cell Res **63**(2): 341-352.

Raghavan, V. (1971). "Synthesis of protein and RNA for initiation and growth of the protonema during germination of bracken fern spore." Exp Cell Res **65**(2): 401-407.

Raghavan, V. (1991). "Gene activity during germination of spores of the fern, Onoclea sensibilis: RNA and protein synthesis and the role of stored mRNA." J Exp Bot **42**(235): 251-260.

Raghavan, V. and H. F. Tung (1967). "Inhibition of two-dimensional growth and suppression of ribonucleic acid and protein synthesis in the gametophytes of the fern, Asplenium nidus, by chloramphenicol, puromycin and actinomycin D." Am J Bot **54**(2): 198-204.

Rau, A., M. Gallopin, G. Celeux and F. Jaffrezic (2013). "Data-based filtering for replicated high-throughput transcriptome sequencing experiments." Bioinformatics **29**(17): 2146-2152.

Renzaglia, K. S., K. D. Wood, G. Rupp and L. G. Hickok (2004). "Characterization of the sleepy sperm mutant in the fern Ceratopteris richardii: A new model for the study of axonemal function." Can. J. Bot. **82**(11): 1602-1617.

Robertson, G., J. Schein, R. Chiu, R. Corbett, M. Field, S. D. Jackman, K. Mungall, S. Lee, H. M. Okada, J. Q. Qian, M. Griffith, A. Raymond, N. Thiessen, T. Cezard, Y. S. Butterfield, R. Newsome, S. K. Chan, R. She, R. Varhol, B. Kamoh, A. L. Prabhu, A. Tam, Y. Zhao, R. A. Moore, M. Hirst, M. A. Marra, S. J. Jones, P. A. Hoodless and I. Birol (2010). "De novo assembly and analysis of RNA-seq data." Nat Methods **7**(11): 909-912.

Robinson, M. D., D. J. McCarthy and G. K. Smyth (2010). "edgeR: a Bioconductor package for differential expression analysis of digital gene expression data." Bioinformatics **26**(1): 139-140.

Robinson, R. W., H. M. Munger, T. W. Whitaker and G. M. Bohn (1976). "Genes of the Cucubitaceae." Hortscience **11**: 554-568.

Rutherford, G., M. Tanurdzic, M. Hasebe and J. A. Banks (2004). "A systemic gene silencing method suitable for high throughput, reverse genetic analyses of gene function in fern gametophytes." BMC plant biology **4**: 6.

Saleh, A., R. Alvarez-Venegas, M. Yilmaz, O. Le, G. Hou, M. Sadder, A. Al-Abdallat, Y. Xia, G. Lu, I. Ladunga and Z. Avramova (2008). "The highly similar Arabidopsis homologs of trithorax ATX1 and ATX2 encode proteins with divergent biochemical functions." Plant Cell **20**(3): 568-579.

Salmi, M. L., T. J. Bushart, S. C. Stout and S. J. Roux (2005). "Profile and analysis of gene expression changes during early development in germinating spores of Ceratopteris richardii." Plant physiology **138**(3): 1734-1745.

Santner, A. and M. Estelle (2010). "The ubiquitin-proteasome system regulates plant hormone signaling." Plant J **61**(6): 1029-1040.

Saze, H., O. Mittelsten Scheid and J. Paszkowski (2003). "Maintenance of CpG methylation is essential for epigenetic inheritance during plant gametogenesis." Nat Genet **34**(1): 65-69.

Schaefer, D. G. and J. P. Zryd (2001). "The moss Physcomitrella patens, now and then." Plant Physiol **127**(4): 1430-1438.

Schmieder, R. and R. Edwards (2011). "Fast Identification and Removal of Sequence Contamination from Genomic and Metagenomic Datasets." Plos One **6**(3).

Schmieder, R. and R. Edwards (2011). "Fast identification and removal of sequence contamination from genomic and metagenomic datasets." PloS one **6**(3): e17288.

Schmieder, R., Y. W. Lim and R. Edwards (2012). "Identification and removal of ribosomal RNA sequences from metatranscriptomes." Bioinformatics **28**(3): 433-435.

Schnable, P. S., F. Hochholdinger and M. Nakazono (2004). "Global expression profiling applied to plant development." Curr Opin Plant Biol **7**(1): 50-56.

Schomburg, D. and I. Schomburg (2001). Springer handbook of enzymes. Berlin ; New York, Springer.

Schuettengruber, B., D. Chourrout, M. Vervoort, B. Leblanc and G. Cavalli (2007). "Genome regulation by polycomb and trithorax proteins." Cell **128**(4): 735-745.

Schulz, M. H., D. R. Zerbino, M. Vingron and E. Birney (2012). "Oases: robust de novo RNA-seq assembly across the dynamic range of expression levels." Bioinformatics **28**(8): 1086-1092.

Scott, R. J. and L. G. Hickok (1991). "Inheritance and Characterization of a Dark-Germinating, Light-Inhibited Mutant in the Fern Ceratopteris-Richardii." Canadian Journal of Botany-Revue Canadienne De Botanique **69**(12): 2616-2619.

Shabek, N. and N. Zheng (2014). "Plant ubiquitin ligases as signaling hubs." Nat Struct Mol Biol **21**(4): 293-296.

Sheard, L. B. and N. Zheng (2009). "Plant biology: Signal advance for abscisic acid." Nature **462**(7273): 575-576.

Slotkin, R. K., M. Vaughn, F. Borges, M. Tanurdzic, J. D. Becker, J. A. Feijo and R. A. Martienssen (2009). "Epigenetic reprogramming and small RNA silencing of transposable elements in pollen." Cell **136**(3): 461-472.

Smith, G. M. (1955). Bryophytes and Pteridophytes. New York, McGraw-Hill.

Smith, S. M. and J. Li (2014). "Signalling and responses to strigolactones and karrikins." Curr Opin Plant Biol **21C**: 23-29.

Strain, E., B. Hass and J. A. Banks (2001). "Characterization of mutations that feminize gametophytes of the fern Ceratopteris." Genetics **159**(3): 1271-1281.

Sun, T. P. (2011). "The molecular mechanism and evolution of the GA-GID1-DELLA signaling module in plants." Curr Biol **21**(9): R338-345.

Sun, T. P. and Y. Kamiya (1994). "The Arabidopsis GA1 locus encodes the cyclase ent-kaurene synthetase A of gibberellin biosynthesis." Plant Cell **6**(10): 1509-1518.

Sussex, I. (1966). "The origin and development of heterospory in vascular plants." Trends in Plant Morphogenesis.

Takeno, K., H. Yamane, T. Yamauchi, N. Takahashi, M. Furber and L. Mander (1989). "Biological activities of the methyl ester of gibberellin a73, a novel and principal antheridiogen in Lygodium japonicum." Plant Cell Physiol. **30**: 201-215.

Tanaka, J., K. Yano, K. Aya, K. Hirano, S. Takehara, E. Koketsu, R. L. Ordonio, S. H. Park, M. Nakajima, M. Ueguchi-Tanaka and M. Matsuoka (2014). "Antheridiogen determines sex in ferns via a spatiotemporally split gibberellin synthesis pathway." Science **346**(6208): 469-473.

Tanurdzic, M. and J. A. Banks (2004). "Sex-determining mechanisms in land plants." Plant Cell **16 Suppl**: S61-71.

Vaughn, K. C., L. G. Hickok, T. R. Warne and A. C. Farrow (1990). "Structural-Analysis and Inheritance of a Clumped-Chloroplast Mutant in the Fern Ceratopteris." Journal of Heredity **81**(2): 146-151.

Wagner, D. and E. M. Meyerowitz (2002). "SPLAYED, a novel SWI/SNF ATPase homolog, controls reproductive development in Arabidopsis." Curr Biol **12**(2): 85-94.

Wald, A. (1943). Tests of statistical hypotheses concerning several parameters when the number of observations is large. *T Amer Math Soc*, *54*, 426-482.

Walker, T. G. (1962). "Cytology and evolution in the fern genus Pteris L." Evolution **16**: 27-43.

Walker, T. G. (1979). The cytogenetics of ferns. The Experimental Biology of Ferns. A. F. Dyer. London, Academic Press**:** 87-132.

Wang, S. S., F. Wang, S. J. Tan, M. X. Wang, N. Sui and X. S. Zhang (2014). "Transcript profiles of maize embryo sacs and preliminary identification of genes involved in the embryo sac-pollen tube interaction." Front Plant Sci **5**: 702.

Wang, Y., X. Zeng, N. J. Iyer, D. W. Bryant, T. C. Mockler and R. Mahalingam (2012). "Exploring the switchgrass transcriptome using second-generation sequencing technology." PLoS One **7**(3): e34225.

Warne, T. R. and L. G. Hickok (1986). "Selection and Characterization of Sodium-Chloride Tolerant Mutants in the Fern Ceratopteris-Richardii." American Journal of Botany **73**(5): 741-741.

Warne, T. R. and L. G. Hickok (1989). "Evidence for a gibberellin biosynthetic origin of ceratopteris antheridiogen." Plant physiology **89**(2): 535-538.

Warne, T. R. and L. G. Hickok (1991). "Control of sexual development in gametophytes of Ceratopteris richardii: Antheridiogen and abscisic acid." Bot. Gaz. **152**: 148-153.

Warne, T. R., L. G. Hickok and R. J. Scott (1988). "Characterization and Genetic-Analysis of Antheridiogen-Insensitive Mutants in the Fern Ceratopteris." Botanical Journal of the Linnean Society **96**(4): 371-379.

Westergaard, M. (1940). "Studies on cytology and sex determination in polyploid forms of *Melandrium album*." Dansk botanisk arkiv **5**: 1-131.

Westergaard, M. (1946). "Aberrant Y chromosomes and sex expression in *Melandrium album*." Hereditas **32**: 419-443.

White, R. A. (1979). Sporophyte development. The Experimental Biology of Ferns. A. F. Dyer. New York, Academic Press.

Whittier, D. P. and T. A. Steeves (1962). "Further studies on induced apogamy in ferns." Can. J. Bot. **40**: 1525-1531.

Wu, X., S. Knapp, A. Stamp, D. K. Stammers, H. Jornvall, S. L. Dellaporta and U. Oppermann (2007). "Biochemical characterization of TASSELSEED 2, an essential plant short-chain dehydrogenase/reductase with broad spectrum activities." FEBS J **274**(5): 1172-1182.

Xi, W., C. Liu, X. Hou and H. Yu (2010). "MOTHER OF FT AND TFL1 regulates seed germination through a negative feedback loop modulating ABA signaling in Arabidopsis." Plant Cell **22**(6): 1733-1748.

Xu, H., Y. Gao and J. Wang (2012). "Transcriptomic analysis of rice (Oryza sativa) developing embryos using the RNA-Seq technique." PLoS One **7**(2): e30646.

Yamane, H. (1998). "Fern Antheridiogens." International Review of Cytology **184**: 1-32.

Yamane, H., K. Nohara, N. Takahashi and H. Schraudolf (1987). "Identification of antheridic acid as an antheridiogen in Anemia rotundifolia and Anemia flexuosa." Plant Cell. Physiol. **28**: 1203-1207.

Yamane, H., Y. Satoh, K. Nohara, M. Nakayama, N. Murofushi, N. Takahashi, K. Takeno, M. Furuya, M. Furber and L. N. Mander (1988). "The methyl ester of a new gibberellin, GA73: the principal antheridiogen in Lygodium japonicum." Tetrahedron Letters **29**: 3959-3962.

Yamane, H., N. Takahashi, K. Takeno and M. Furuya (1979). "Identification of gibberellin A9 methyl ester as a natural substance regulating formation of reproductive organs in Lygodium japonicum." Planta **147**: 251-256.

Yamasaki, S., N. Fujii, S. Matsuura, H. Mizusawa and H. Takahashi (2001). "The M locus and ethylene-controlled sex determination in andromonoecious cucumber plants." Plant Cell Physiol **42**(6): 608-619.

Yamasaki, S., N. Fujii and H. Takahashi (2005). "Hormonal regulation of sex expression in plants." Vitamins and hormones **72**: 79.

Yang, W., M. Pollard, Y. Li-Beisson, F. Beisson, M. Feig and J. Ohlrogge (2010). "A distinct type of glycerol-3-phosphate acyltransferase with sn-2 preference and phosphatase activity producing 2-monoacylglycerol." Proc Natl Acad Sci U S A **107**(26): 12040-12045.

Ye, D., P. Installé, C. Ciuperescu, J. Veuskens, Y. Wu, G. Salesses, M. Jacobs and I. Negrutiu (1990). " Sex determination in the dioecious *Melandrium*. I. First lessons from androgenic haploids." Sexual Plant Reproduction **3**: 179-186.

Yin, T. and J. A. Quinn (1995). "Tests of a mechanistic model of one hormone regulating both sexes in *Cucumis sativus* (Cucurbitaceae)." American Journal of Botany **82**(12): 1537-1546.

Young, M. D., M. J. Wakefield, G. K. Smyth and A. Oshlack (2010). "Gene ontology analysis for RNA-seq: accounting for selection bias." Genome Biol **11**(2): R14.

Zdobnov, E. M. and R. Apweiler (2001). "InterProScan--an integration platform for the signature-recognition methods in InterPro." Bioinformatics **17**(9): 847-848.

Zenoni, S., A. Ferrarini, E. Giacomelli, L. Xumerle, M. Fasoli, G. Malerba, D. Bellin, M. Pezzotti and M. Delledonne (2010). "Characterization of transcriptional complexity during berry development in Vitis vinifera using RNA-Seq." Plant Physiol **152**(4): 1787-1795.

Zhang, J., T. A. Ruhlman, J. P. Mower and R. K. Jansen (2013). "Comparative analyses of two Geraniaceae transcriptomes using next-generation sequencing." BMC Plant Biol **13**: 228.

Zhang, W., J. P. To, C. Y. Cheng, G. E. Schaller and J. J. Kieber (2011). "Type-A response regulators are required for proper root apical meristem function through post-transcriptional regulation of PIN auxin efflux carriers." Plant J **68**(1): 1-10.

Zhang, Y., X. Zhang, B. Liu, W. Wang, X. Liu, C. Chen, X. Liu, S. Yang and H. Ren (2014). "A GAMYB homologue CsGAMYB1 regulates sex expression of cucumber via an ethylene-independent pathway." J Exp Bot **65**(12): 3201-3213.

Zhu, J., A. Kapoor, V. V. Sridhar, F. Agius and J. K. Zhu (2007). "The DNA glycosylase/lyase ROS1 functions in pruning DNA methylation patterns in Arabidopsis." Curr Biol **17**(1): 54-59.

Zluvova, J., M. Nicolas, A. Berger, I. Negrutiu and F. Moneger (2006). "Premature arrest of the male flower meristem precedes sexual dimorphism in the dioecious plant Silene

APPENDICES

Appendix A Computer Scripts

```
######################CHAPTER 2 SCRIPTS##############################
##Check read quality with FastQC##
fastqc -o /scratch/lustreA/n/natallah/FastQCreports --noextract -f fastq CFM1_1_Trim.fq
CFM1_2_Trim.fq CFM2_1_Trim.fq CFM2_2_Trim.fq CFM3_1_Trim.fq
CFM3_2_Trim.fq FFM1_1_Trim.fq FFM1_2_Trim.fq FFM2_1_Trim.fq
FFM2_2_Trim.fq FFM3_1_Trim.fq FFM3_2_Trim.fq

##An example of Trimmomatic script on one fastq file##
java -classpath /apps/group/bioinformatics/apps/trimmomatic-0.20/trimmomatic-0.20.jar
org.usadellab.trimmomatic.TrimmomaticSE -phred33 -trimlog
FFM2_1.fastq_clean.fq.no_adapter.trimmomatic.trim FFM2_1.fastq_clean.fq.no_adapter
FFM2_1.fastq_clean.fq.no_adapter.trimmomatic LEADING:7 TRAILING:7
SLIDINGWINDOW:4:13 MINLEN:30 >
FFM2_1.fastq_clean.fq.no_adapter.trimmomatic.log

##Deconseq script#
perl deconseq/deconseq-standalone-0.4.1/deconseq.pl -keep_tmp_files -c 50 -i 75 -dbs
rna,wmitochondria,wchloroplast,virus,bacteria  -id CFM2_2.fastq -f CFM2_2.fastq

##For running Trinity, first need to concatenate all cleaned/trimmed reads into left reads
and into right reads.
#To concatenate left reads:
cat CFM1_1.fastq_clean.fq.no_adapter.trimmomatic
CFM2_1.fastq_clean.fq.no_adapter.trimmomatic
CFM3_1.fastq_clean.fq.no_adapter.trimmomatic
FFM1_1.fastq_clean.fq.no_adapter.trimmomatic
FFM2_1.fastq_clean.fq.no_adapter.trimmomatic
FFM3_1.fastq_clean.fq.no_adapter.trimmomatic >clean_left_reads.fastq
#To concatenate right reads:
cat CFM1_2.fastq_clean.fq.no_adapter.trimmomatic
CFM2_2.fastq_clean.fq.no_adapter.trimmomatic
CFM3_2.fastq_clean.fq.no_adapter.trimmomatic
FFM1_2.fastq_clean.fq.no_adapter.trimmomatic
FFM2_2.fastq_clean.fq.no_adapter.trimmomatic
FFM3_2.fastq_clean.fq.no_adapter.trimmomatic >clean_right_reads.fastq

##get number of reads by doing (depending on what the fastq headers are like)##
grep -c "^@ILLUMINA" CFM1_2_Trim.fq
#or
grep -c "^@HW-ST994" CFM3_1_Trim.fq
```

```
##run trinity##
trinityrnaseq_r2012-06-08/Trinity.pl --seqType fq–JM 100G --left clean_left_reads.fastq
--right clean_right_reads.fastq --output trinityout150 --min_contig_length 150 --CPU 24 -
-bfly_opts “--bflyCPU 24"

##example RSEM commands##
extract-transcript-to-gene-map-from-trinity Trinity.fasta map_Trinity

rsem-prepare-reference --transcript-to-gene-map map_Trinity
--no-polyA Trinity.fasta referenceTrinity


rsem-calculate-expression   --calc-ci --out-bam --paired-end CFM2_1.fastq
CFM2_2.fastq referenceTrinity CFM2inAll6counts

rsem-bam2wig FFM2counts wig_FFM2 wiggle_FFM2rse

rsem-plot-model CFM2inAll6counts plot_CFM2inAll6model.pdf

rsem-calculate-expression --paired-end --bowtie-chunkmbs 200 --strand-specific -p 8
pairedReads/pairedReads/AP_dry_spores_R1_clean.fq.no_adapter.pair
pairedReads/pairedReads/AP_dry_spores_R2_clean.fq.no_adapter.pair referenceTrinity
AP_dry_spores


##blast Trinity assembly against Selaginella and Arabidopsis proteins##
#make custom database
makeblastdb -in SelmoArab.fasta -dbtype prot
#blastx
blastx -query uniqueCompCleanExp.fasta -out ExpTrinityvsSelmoArab -db
SelmoArab_aa.fasta -evalue 0.0000000001 -outfmt '6 qseqid qlen sseqid slen qstart qend
sstart send length pident bitscore evalue' -show_gis -num_threads 8
#how many unique contigs have hits
cut -f 1 ExpTrinityvsSelmoArab | sort | uniq | wc –l
#19217 have hits (23%)

###which sequences have homology with Arabidopsis homoebox leucine zipper family
proteins###
grep
'AT1G34650\|AT1G73360\|AT2G01430\|AT2G32370\|AT3G03260\|AT3G61150\|AT4G
17710\|AT5G06710\|AT5G17320\|AT5G52170\|AT5G47370' TrinityvsSelmoArab | grep
-o 'comp[0-9]*'  | sort| uniq

##All against all blast##
```

```
#make custom database from Trinity assembly
makeblastdb -in Trinity.fasta.tmp -dbtype nucl
#blastn
blastn -query Trinity.fasta.tmp -out TrinityvsTrinity -db Trinity.fasta.tmp -evalue
0.0000000001 -outfmt '6 qseqid qlen sseqid slen qstart qend sstart send length pident
bitscore evalue' -show_gis -num_threads 8

##make histogram of blast hits' bitscores in R##
#for blasting CrESTs in Genbank against Trinity assembly
genbank<-read.table("GenbankCrESTsvsCleanTrinBlast")
hist(genbank$V11, xlab= "Bitscore", ylab= "Number of Sequences", main= "Disribution
of Bitscores obtained with BLASTn of Genbank Ceratopteris ESTs against Ceratopteris
Trinity Assembly", col="red", labels=TRUE, ylim=c(0,2500), xlim=c(0,2500))


###get A. thaliana accessions for genes up in male or genes up in hermaphrodite to do
enrichment test on in AgriGO###
#first copy and paste genes names and A. thaliana accessions for all DEGs into
spreadsheet.  Leave only _seq1's
#so that we don't bias enrichment test towards genes with multiple isoforms. Do this in
Unix:
grep "_seq1len" AllseqsnamesAthalMatch.txt > f
#Remove duplicate lines now
sort play | uniq -u > f2
#sort DE results in excel based on DESeq fold change to separate components up in M vs
H and then in Unix for male and her files do:
join <(sort f1) <(sort f2)

Remove duplicate lines so as to only
### check assembly quality####
#blast with blastn version 2.2.28+
blastn -query CrESTS5000 -out GenbankCrESTsvsCleanTrinBlast -db CleanCrContigs -
evalue 0.0000000001 -outfmt '6 qseqid qlen sseqid slen qstart qend sstart send length
pident bitscore evalue' -show_gis -num_threads 8

grep -c gi CrESTS5000

cut -f 1  GenbankCrESTsvsCleanTrinBlast | sort | uniq | wc –l

#####R commands (general)#####
w<-read.table("FFM1comps.genes.results")
w1<-read.table("FFM2comps.genes.results")
w2<-read.table("FFM3comps.genes.results")
x<-read.table("CFM1comps.genes.results")
x1<-read.table("CFM2comps.genes.results")
```

```
x2<-read.table("CFM3comps.genes.results")
counts=matrix(0,dim(x)[1],6)
counts[,1]=as.integer(w$V2)
counts[,2]=as.integer(w1$V2)
counts[,3]=as.integer(w2$V2)
counts[,4]=as.integer(x$V2)
counts[,5]=as.integer(x1$V2)
counts[,6]=as.integer(x2$V2)
colnames(counts)=c('-ACE1','-ACE2','-ACE3','+ACE1','+ACE2','+ACE3')
rownames(counts)=x$V1
counts=counts[rowSums(counts)!=0,]


##edgeR commands##
library(edgeR)
conds= c(rep("-ACE",3),rep("+ACE",3))
#make data object
cds = DGEList(counts, group=conds)
#normalizes by finding scaling factors for library sizes that minimize the log-FC between
samples (TMM)
cds <- calcNormFactors(cds)
cds$samples$lib.size * cds$samples$norm.factors
cds <- estimateTagwiseDisp(cds)

de.tgw = exactTest(cds,dispersion='tagwise',pair=c("-ACE","+ACE"))
de.tgw$table$logFC.abs=abs(de.tgw$table$logFC)
sum(p.adjust(de.tgw$table$PValue, method = "BH") < 0.01)
deg.tgw = de.tgw[(p.adjust(de.tgw$table$PValue, method = "BH") < 0.01),]
fc2 = deg.tgw[which(deg.tgw$table$logFC.abs>1),]
dim(fc2)
sum(fc2$table$logFC<0)
sum(fc2$table$logFC>0)
write.csv(fc2$table,file='edgeRcompFC2')
156

##DESeq commands##
library(DESeq)
#make data structure
decds<-newCountDataSet( counts,conds )
head(counts(decds))
#estimate effective library size
decds<-estimateSizeFactors(decds)
#estimate dispersion (BCV2)
decds<-estimateDispersions(decds)
```

```
#list fit info object and structure (contains values used in inference that result from prior
step)
str( fitInfo(decds))
#negative binomial test to check for differential expression
res<-nbinomTest(decds, "-ACE", "+ACE" )
head(res)
dim(res)
res$logFC.abs=abs(res$log2FoldChange)
sum(res$padj < 0.01)
de.2 <- res[ res$padj < 0.01, ]
de.2=de.2[de.2$logFC.abs>1,]
sum(de.2$log2FoldChange<0)
sum(de.2$log2FoldChange>0)
157

##EBSeq commands##
#load EBSeq and necessary packages into working space
library(blockmodeling,lib.loc=".")
library(EBSeq,lib.loc=".")
library(hexbin)
library(latticeExtra)
library(gplots)
library(geneplotter)
#estimate size factors in same manner as DESeq
Sizes = MedianNorm(counts)
#look for DEGs
EBOut = EBTest(Data = counts, Conditions = as.factor(rep(c("-
ACE","+ACE"),each=3)),sizeFactors = Sizes, maxround = 10)
PP=GetPPMat(EBOut)   #gets a matrix of the posterior probabilities
par(mfrow=c(2,2))
QQP(EBOut)
DenNHist(EBOut)
p.adjust(PP[,"PPEE"], method = "BH")
DEfound = rownames(PP)[which(PP[, "PPDE"] > 0.99)]
c1=unlist(EBOut$C1Mean)          # vector of mean expression in FM
c2=unlist(EBOut$C2Mean)          # vector of mean expression in CFM
c1.de=c1[DEfound]
c2.de=c2[DEfound]
logfc=log(c2/c1,base=2)          # compute log fold change
sum(logfc[DEfound]>0)            # number of upregulated genes
DEfound.2fc=names(logfc[DEfound][abs(logfc[DEfound])>1])
157
#get distribution of average normalized counts per gene#
counts<-read.csv("NormalizedCountsAllGenes.csv")
normcounts<-as.matrix(counts[,2:7])
```

```
rownames(normcounts)=counts$X
hist(rowMeans(normcounts),xlab="Mean Read Depth",
ylab="Frequency",main="Distribution of Read Depth Across
Components",col="green",labels=FALSE,xlim=c(0,5000), breaks=700, ylim=c(0,6000))
box(which = "plot", lty="solid")
# plot baseMeans against each other
plot(log2(res$baseMeanA),log2(res$baseMeanB), pch=".", cex=.3, ylab="log2(baseMean)
+ACE", xlab="log2(baseMean) -ACE", col=ifelse(res$padj<0.01, "red","black"))

####GO enrichment test##################
library(goseq)
library(GO.db)
library("biomaRt")

#median of isoform lengths
lengthData<-read.table("CompsAndMedLen.txt",row.names=1)

#go annotation using blast results against blastx
#format: comp10000<TAB>GO:1919191, one comp-go pair a line
go <- read.table("AllCleanContigsExpGOformatted.txt", header=FALSE, sep="\t",
fill=TRUE)
head(go)
#get GOslim terms from BioMart
ensembl <- useMart("plants_mart_23",dataset="athaliana_eg_gene")
go_slim<-getBM(attributes="goslim_goa_accession",mart=ensembl)[,1]
#filter GO terms to keep only GOslim terms
go_slim2cat<-subset(go, go[,2] %in% go_slim)
#names of all comp names kept in DEG analysis
keep <- read.table('allgenesNames.txt')

#all DEGs identified
male.genes<-read.table("IDs_1162Male.txt")
herm.genes<-read.table("IDs_1162Herm.txt")
DEG.genes<-read.table("IDs_1162DEGs.txt")

Mgenes=as.integer(keep[,1]%in%male.genes[,1])
names(Mgenes)=keep[,1]
head(Mgenes)

Hgenes=as.integer(keep[,1]%in%her.genes[,1])
names(Hgenes)=keep[,1]
head(Hgenes)

Mbias=lengthData[rownames(lengthData)%in%names(Mgenes),]
names(Mbias) = rownames(lengthData)[rownames(lengthData)%in%names(Mgenes)]
```

```
head(Mbias)

Hbias=lengthData[rownames(lengthData)%in%names(Hgenes),]
names(Hbias) = rownames(lengthData)[rownames(lengthData)%in%names(Hgenes)]
head(Hbias)


Mpwf = nullp(Mgenes,bias.data=Mbias)
Hpwf = nullp(Hgenes,bias.data=Hbias)

GO.wall.M <- goseq(Hpwf, gene2cat=go
GO.wall.H <- goseq(Hpwf, gene2cat=go)
GO.wall.M=goseq(Mpwf,gene2cat=go_slim2cat)
GO.wall.H=goseq(Hpwf,gene2cat=go_slim2cat)
head(GO.wall.M)
enriched.GO.M = GO.wall.M$category[GO.wall.M$over_represented_pvalue <=0.05]
enriched.GO.H = GO.wall.H$category[GO.wall.H$over_represented_pvalue <=0.05]
head(enriched.GO.M)
#print into  file
sink(file="enrichedGOannot_maleGOslim0.1.txt")
for(go in enriched.GO.M[1:length(enriched.GO.M)]){print(GOTERM[[go]])
cat("--------------------------------------\n")
}
sink()


##run RepeatMasker###
RepeatMasker  -species viridiplantae -gccalc  uniqueCompCleanExp.fasta

##Trinotate##
transcripts_to_best_scoring_ORFs.pl -t uniqueCompCleanExp.fasta -m 50

ftp://ftp.uniprot.org/pub/databases/uniprot/current_release/knowledgebase/complete/unipr
ot_sprot.fasta.gz
blastp -query best_candidates.eclipsed_orfs_removed.pep -db SwissProtFormated -
num_threads 8 -max_target_seqs 1 -outfmt 6 -out TrinotateBlast.out
makeblastdb -in uniprot_sprot.fasta -dbtype prot

blastp -query best_candidates.eclipsed_orfs_removed.pep -db uniprot_sprot.fasta -evalue
0.0000000001 -num_threads 8 -max_target_seqs 1 -outfmt 6 -out TrinotateBlast.out
ftp://ftp.sanger.ac.uk/pub/databases/Pfam/current_release/Pfam-A.hmm.gz

hmmpress Pfam-A.hmm
```

```
hmmscan --cpu 8 --domtblout TrinotatePFAM.out Pfam-A.hmm
best_candidates.eclipsed_orfs_removed.pep > pfam.log

signalp -f short -n signalp.out best_candidates.eclipsed_orfs_removed.pep

tmhmm --short < best_candidates.eclipsed_orfs_removed.pep > tmhmm.out

###
##Downloaded the Trinity sqlite database with swissprot-related info from :
##http://sourceforge.net/projects/trinityrnaseq/files/misc/TRINOTATE_RESOURCES/Tr
inityFunctional.swissprot.2012-02-13.db.gz/download
###

Trinotate.pl LOAD_transdecoder best_candidates.eclipsed_orfs_removed.pep
Trinotate.pl LOAD_blast TrinotateBlast.out
Trinotate.pl LOAD_pfam TrinotatePFAM.out
Trinotate.pl LOAD_signalp signalp.out
Trinotate.pl LOAD_tmhmm tmhmm.out
Trinotate.pl report -E 0.0000000001 > trinotate_annotation_report.xls

##see how many unique sequences have ORFs greater than the cutoff##
grep -o 'comp[0-9]*_c[0-9]*_seq[0-9]*' best_candidates.eclipsed_orfs_removed.pep |
sort  | uniq | wc -l


########Custom Perl
Scripts#############################################################################
########
#############################################################################
##################
#                   getComponentMedianLen.pl
#
#   Takes as input a file with trinity components and lengths and outputs the median
length
#         for each component
#         input file should be text with: component\tlength
#
#         getComponentMedianLen.pl inFile > outfile.txt
#
#                   Written by Nadia Atallah  on 28 Oct 2014
#
#############################################################################
##################

#-----------------
```

```
# Begin Script
#-----------------

#!/usr/bin/perl

use strict;
use warnings;
use Statistics::Descriptive;



my $stat = Statistics::Descriptive::Full->new();
my @data=();
my ( $newName, $oldName, $line );
my $i=0;

#read lines in
while ( $line = <> ) {
	chomp $line;
	if ( $i > 0 ) { $oldName=$newName; }	#keep track of both new and old names
for comparison
	$i++;
	my ($name,$len) = split " ", $line,2;
	$newName=$name;
	if( $i==1) {
		push @data,$len;
	} elsif ( $newName eq $oldName ) {
		push @data,$len;
	} else {
		my $stat = Statistics::Descriptive::Full->new();
		$stat->add_data(@data);
		print "$oldName\t".$stat->median() . "\n";
		@data=();
		push @data,$len;
	}
}
my $stat = Statistics::Descriptive::Full->new();
$stat->add_data(@data);
print "$newName\t".$stat->median() . "\n";

#-----------------
# End Script
#-----------------
```

```
##############################################################################
##################
#this program takes as input a Trinity fasta file and outputs a file with the names,
#	documentation, and sequences of the desired genes
#
#Nadia Atallah		12 march, 2012
## run it with this perl script:  getDESequences.pl inputfile > outputfile
##############################################################################
##################

#!/usr/bin/perl

#-----------------
# Begin Script
#-----------------

use strict;
#make an array of the names of DE genes
my @lookfor = qw(
#####put gene names in here#######
);

my ( $line, $name, $doc);
my $currentbases = "";
my $Is_Good = 0;			# Indicator of whether current sequence $name is
good
my $Prev_Was_Good = 0;		# Indicator of whether previous sequence $name
was good

while ( $line = <> ) {						#read lines in
	chomp $line;							#remove end of line
character
	if ( $line =~ /^>/ ) {					#check if line begins with >
		if ( $Is_Good == 1 ) {  print $currentbases; print "\n";}	# If previous
sequence was good print its bases
		( $name, $doc ) = split " ",  $line, 2;		#extract the name and
documentation of the sequence
		$name =~ s/>//;			#get rid of >
		$Is_Good = 0;
		foreach my $j ( 0 .. $#lookfor) {
			if (  $name =~ /^$lookfor[$j]$/ ) {$Is_Good = 1;}
		}
		if ( $Is_Good == 1 ) { print ">"; print $name; print "\n";}  # Print name of
sequence if it is good
```

```
                $currentbases = "";                                     #reset the
currentbases string to empty
        }
        else {
                        $currentbases .= $line;                         #add line to sequence
string
                }
        }
        if ( $Is_Good == 1)   {print $currentbases; print "\n";}                #This kicks in
if end of file is encountered and sequence was good

#-----------------
# End Script
#-----------------

##############################################################################
##################
# This program takes as input tablular blast output and prints only the top
# blast hit for each sequence
# Written by Nadia Atallah on 1 October 2013
#
##############################################################################
##################


#!/usr/bin/perl

#-----------------
#Begin Script
#-----------------

use warnings;
use strict;

my ($line, $name1, $old_name);
my @result = ();
my $i=0;

while ( $line = <> ) {
        chomp $line;
        @result = split " ", $line;
        $i++;
        if ( $i > 1 ) {
                $old_name = $name1;
        }
```

```
        $name1 = $result[0];
        if ( $old_name eq $name1 ) {
                next;
        } else { print "$line\n"; }
}

if ( $old_name ne $name1 ) {
        print $line;
 }

#-----------------
# End Script
#-----------------

##############################################################################
####
#       formatGOterms.pl
#
# This program takes as input a text file with one component (or accession)
# on each line then GO terms:
# component\tGOterms
#
#      usage: formatGOterms.pl infile > outfile
#
#          Written by Nadia Atallah on 30 October 2014
#
##############################################################################
####

#-----------------
# Begin Script
#-----------------


#!/usr/bin/perl

use warnings;
use strict;

my @goarray = ();
my ( $goterms, $accession, $line );

while ( $line = <> ) {
        chomp $line;
        ( $accession, $goterms ) = split "\t", $line;
```

```
        @goarray = split ",", $goterms;
        foreach my $i ( 0 .. $#goarray) {
                print "$accession\t$goarray[$i]\n";
        }
}

#-----------------
# End Script
#-----------------


######################CHAPTER 3 SCRIPTS#################################

#Time-course R analysis
source("http://bioconductor.org/biocLite.R")
library('DESeq2')
library('Biobase')
library('DESeq')
library('edgeR')
library('genefilter')
library('gplots')
setwd("~/Desktop")

file_namesCFM<-c('drySpores_1','dry_spores_2','3_1','3_2','3_5',
'3.5_+ACE1','3.5_+ACE2','3.5_+ACE5',
            '4.5_+ACE1','4.5_+ACE2','4.5_+ACE5',
'5.5_+ACE1','5.5_+ACE2','5.5_+ACE5')
time<-c('0dai','0dai','3dai','3dai','3dai', '3.5dai','3.5dai','3.5dai', '4.5dai','4.5dai','4.5dai',
      '5.5dai','5.5dai','5.5dai')
bioRep<-c('1','2','1','2','5','1','2','5','1','2','5','1','2','5')

samplesCFM<-data.frame(file_namesCFM,time)
samplesCFM <-
data.frame(row.names=c("0d1","0d2","3d1","3d2","3d5","3.5d1","3.5d2","3.5d5","4.5d1
","4.5d2","4.5d5",
                        "5.5d1","5.5d2","5.5d5"),
                time=as.factor(c(rep("0d",2),
rep("3d",3),rep("3.5d",3),rep("4.5d",3),rep("5.5d",3))))

spores1<-
read.table("rsemReferenceGenesResults/AP_dry_spores_Reference.genes.results",header
=TRUE)
spores2<-
read.table("rsemReferenceGenesResults/KE_dry_spores_Reference.genes.results",header
=TRUE)
```

```
FM3_1<-
read.table("rsemReferenceGenesResults/1_3daiReference.genes.results",header=TRUE)
FM3_2<-
read.table("rsemReferenceGenesResults/2_3daiReference.genes.results",header=TRUE)
FM3_5<-
read.table("rsemReferenceGenesResults/5_3daiReference.genes.results",header=TRUE)
CFM1_3_5<-
read.table("rsemReferenceGenesResults/CFM1_3_5dai_Reference.genes.results",header
=TRUE)
CFM2_3_5<-
read.table("rsemReferenceGenesResults/CFM2_3_5daiReference.genes.results",header=
TRUE)
CFM5_3_5<-
read.table("rsemReferenceGenesResults/CFM5_3_5daiReference.genes.results",header=
TRUE)
CFM1_4_5<-
read.table("rsemReferenceGenesResults/CFM1_4_5daiReference.genes.results",header=
TRUE)
CFM2_4_5<-
read.table("rsemReferenceGenesResults/CFM2_4_5daiReference.genes.results",header=
TRUE)
CFM5_4_5<-
read.table("rsemReferenceGenesResults/CFM5_4_5daiReference.genes.results",header=
TRUE)
CFM1_5_5<-
read.table("rsemReferenceGenesResults/CFM1_5_5daiReference.genes.results",header=
TRUE)
CFM2_5_5<-
read.table("rsemReferenceGenesResults/CFM2_5_5daiReference.genes.results",header=
TRUE)
CFM5_5_5<-
read.table("rsemReferenceGenesResults/CFM5_5_5daiReference.genes.results",header=
TRUE)

countsCFM<-matrix(0,dim(CFM5_5_5)[1],14)

countszero<-matrix(0,dim(CFM5_5_5)[1],2)
countsrest<-matrix(0,dim(CFM5_5_5)[1],12)

#figure out genes that are stage-specifically expressed
counts0<-matrix(0,dim(CFM5_5_5)[1],2)
counts3<-matrix(0,dim(CFM5_5_5)[1],3)
counts3.5<-matrix(0,dim(CFM5_5_5)[1],3)
counts4.5<-matrix(0,dim(CFM5_5_5)[1],3)
counts5.5<-matrix(0,dim(CFM5_5_5)[1],3)
```

```
counts0[,1]<-as.integer(spores1$expected_count)
counts0[,2]<-as.integer(spores2$expected_count)
counts3[,1]<-as.integer(FM3_1$expected_count)
counts3[,2]<-as.integer(FM3_2$expected_count)
counts3[,3]<-as.integer(FM3_5$expected_count)
counts3.5[,1]<-as.integer(CFM1_3_5$expected_count)
counts3.5[,2]<-as.integer(CFM2_3_5$expected_count)
counts3.5[,3]<-as.integer(CFM5_3_5$expected_count)
counts4.5[,1]<-as.integer(CFM1_4_5$expected_count)
counts4.5[,2]<-as.integer(CFM2_4_5$expected_count)
counts4.5[,3]<-as.integer(CFM5_4_5$expected_count)
counts5.5[,1]<-as.integer(CFM1_5_5$expected_count)
counts5.5[,2]<-as.integer(CFM2_5_5$expected_count)
counts5.5[,3]<-as.integer(CFM5_5_5$expected_count)
rownames(counts0)=CFM5_5_5$gene_id
rownames(counts3)=CFM5_5_5$gene_id
rownames(counts3.5)=CFM5_5_5$gene_id
rownames(counts4.5)=CFM5_5_5$gene_id
rownames(counts5.5)=CFM5_5_5$gene_id
counts0=counts0[rowSums(counts0)!=0,]
counts3=counts3[rowSums(counts3)!=0,]
counts3.5=counts3.5[rowSums(counts3.5)!=0,]
counts4.5=counts4.5[rowSums(counts4.5)!=0,]
counts5.5=counts5.5[rowSums(counts5.5)!=0,]
colnames(counts0)=c('0dai1','0dai2')
colnames(counts3)=c('3dai1','3dai2','3dai5')
colnames(counts3.5)=c('+ACE3.5dai1','+ACE3.5dai2','+ACE3.5dai5')
colnames(counts4.5)=c( '+ACE4.5dai1','+ACE4.5dai2','+ACE4.5dai5')
colnames(counts5.5)=c('+ACE5.5dai1','+ACE5.5dai2','+ACE5.5dai5')
cpm.3 <- counts3[rowSums(1e+06 * counts3/expandAsMatrix(colSums(counts3),
dim(counts3)) > 0.3) >= 3, ]
cpm.0<-counts0[rowSums(1e+06 * counts0/expandAsMatrix(colSums(counts0),
dim(counts0)) > 0.3) >= 2, ]
cpm.3.5 <- counts3.5[rowSums(1e+06 * counts3.5/expandAsMatrix(colSums(counts3.5),
dim(counts3.5)) > 0.3) >= 3, ]
cpm.4.5 <- counts4.5[rowSums(1e+06 * counts4.5/expandAsMatrix(colSums(counts4.5),
dim(counts4.5)) > 0.3) >= 3, ]
cpm.5.5 <- counts5.5[rowSums(1e+06 * counts5.5/expandAsMatrix(colSums(counts5.5),
dim(counts5.5)) > 0.3) >= 3, ]
library('gplots')
boom <-
list("0DAI"=rownames(cpm.0),"3DAI"=rownames(cpm.3),"3.5DAI"=rownames(cpm.3.5
),"4.5DAI"=rownames(cpm.4.5),"5.5DAI"=rownames(cpm.5.5))
venn(boom)
```

```
#get only genes that are expressed specifically at the 0 day time-point



countsCFM[,1]<-as.integer(spores1$expected_count)
countsCFM[,2]<-as.integer(spores2$expected_count)
countsCFM[,3]<-as.integer(FM3_1$expected_count)
countsCFM[,4]<-as.integer(FM3_2$expected_count)
countsCFM[,5]<-as.integer(FM3_5$expected_count)
countsCFM[,6]<-as.integer(CFM1_3_5$expected_count)
countsCFM[,7]<-as.integer(CFM2_3_5$expected_count)
countsCFM[,8]<-as.integer(CFM5_3_5$expected_count)
countsCFM[,9]<-as.integer(CFM1_4_5$expected_count)
countsCFM[,10]<-as.integer(CFM2_4_5$expected_count)
countsCFM[,11]<-as.integer(CFM5_4_5$expected_count)
countsCFM[,12]<-as.integer(CFM1_5_5$expected_count)
countsCFM[,13]<-as.integer(CFM2_5_5$expected_count)
countsCFM[,14]<-as.integer(CFM5_5_5$expected_count)


countszero[,1]<-countsCFM[,1]
countszero[,2]<-countsCFM[,2]

countsrest[,1]<-countsCFM[,3]
countsrest[,2]<-countsCFM[,4]
countsrest[,3]<-countsCFM[,5]
countsrest[,4]<-countsCFM[,6]
countsrest[,5]<-countsCFM[,7]
countsrest[,6]<-countsCFM[,8]
countsrest[,7]<-countsCFM[,9]
countsrest[,8]<-countsCFM[,10]
countsrest[,9]<-countsCFM[,11]
countsrest[,10]<-countsCFM[,12]
countsrest[,11]<-countsCFM[,13]
countsrest[,12]<-countsCFM[,14]

rownames(countsCFM)=CFM5_5_5$gene_id
rownames(countszero)=CFM5_5_5$gene_id
rownames(countsrest)=CFM5_5_5$gene_id

countszero=countszero[rowSums(countszero)!=0,]
countsrest=countsrest[rowSums(countsrest)!=0,]
```

```
countsCFM=countsCFM[rowSums(countsCFM)!=0,]

colnames(countsCFM)=c('0dai1','0dai2',
                '3dai1','3dai2','3dai5',
                '+ACE3.5dai1','+ACE3.5dai2','+ACE3.5dai5',
                '+ACE4.5dai1','+ACE4.5dai2','+ACE4.5dai5',
                '+ACE5.5dai1','+ACE5.5dai2','+ACE5.5dai5'
)
colnames(countszero)=c('0dai1','0dai2')
colnames(countsrest)=c('3dai1','3dai2','3dai5',
                 '+ACE3.5dai1','+ACE3.5dai2','+ACE3.5dai5',
                 '+ACE4.5dai1','+ACE4.5dai2','+ACE4.5dai5',
                 '+ACE5.5dai1','+ACE5.5dai2','+ACE5.5dai5'
)

#filter everything under 0.3CPM, keeping in mind that 0 day time-point has only 2
replicates
cpm.3 <- countsrest[rowSums(1e+06 * countsrest/expandAsMatrix(colSums(countsrest),
dim(countsrest)) > 0.3) >= 3, ]
cpm.2<-countszero[rowSums(1e+06 * countszero/expandAsMatrix(colSums(countszero),
dim(countszero)) > 0.3) >= 2, ]

library('sets')
name3<-rownames(cpm.3)
name2<-rownames(cpm.2)
union<-union(name3,name2)
keep3<-countsCFM[union,]
length(union)
dim(keep3)
colnames(keep3)=c('0dai1','0dai2',
                '3dai1','3dai2','3dai5',
                '+ACE3.5dai1','+ACE3.5dai2','+ACE3.5dai5',
                '+ACE4.5dai1','+ACE4.5dai2','+ACE4.5dai5',
                '+ACE5.5dai1','+ACE5.5dai2','+ACE5.5dai5'
)
#write.csv(keep3,file="rawCountsPassingFiltering")
#############differential expression analysis###################
library('DESeq')
dataCFM<-newCountDataSet(keep3,samplesCFM)
countDataCFM<-counts(dataCFM)
colDataCFM<-pData(dataCFM)[,"time"]
ddsCFM <- DESeqDataSetFromMatrix(countData = countDataCFM,
                        colData = samplesCFM,
                        design = ~ time)
```

```
ddsCFM <- DESeq(ddsCFM,betaPrior=FALSE)
#do a log ratio test
#the null is that there is no condition effect and the same time effect for all conditions

ddsLRT <- nbinomLRT(ddsCFM, reduced = ~ 1)
resLRT <- results(ddsLRT,independentFiltering=FALSE)
resLRT <-na.omit(resLRT)

resLRT <- resLRT[order(resLRT$padj),]
head(resLRT)
resLRT$logFC.abs<-abs(resLRT$log2FoldChange)
sum(resLRT$padj < 0.05)
degLRT <- resLRT[resLRT$padj < 0.05, ]
fc2_LRT = degLRT[which(degLRT$logFC.abs>1),]
dim(fc2_LRT)
sum(fc2_LRT$log2FoldChange<0)
sum(fc2_LRT$log2FoldChange>0)
#####compare 0-3, 3-3.5, 3.5-4.5,4.5-5.5#########
###0-3###
colData(ddsCFM)$time <- relevel(colData(ddsCFM)$time, "0d")
ddsCFM <- DESeq(ddsCFM,betaPrior=FALSE)
resCFM <- results(ddsCFM)
resultsNames(ddsCFM)
mcols(resCFM, use.names=TRUE)
res0CFM <- results(ddsCFM,"time_3d_vs_0d")
res0CFMna <- results(ddsCFM,"time_3d_vs_0d")
res0CFM <-na.omit(res0CFM)
sum(res0CFM$padj < 0.05)
deg0CFM <- res0CFM[res0CFM$padj < 0.05, ]
deg0CFM$logFC.abs=abs(deg0CFM$log2FoldChange)
fc2_0 = deg0CFM[which(deg0CFM$logFC.abs>1),]
dim(fc2_0)
#write.csv(deg0CFM,file="DEGs0-3d")
plot(res0CFM$log2FoldChange, -log10(res0CFM$padj),pch=".",main="Volcano Plot for
0-3DAI Samples",col=ifelse(res0CFM$padj<0.05, "red","black"),xlab="log2(Fold
Change)",ylab="-log10(padj)")
sum(fc2_0$log2FoldChange<0)
sum(fc2_0$log2FoldChange>0)
###3-3.5###
colData(ddsCFM)$time <- relevel(colData(ddsCFM)$time, "3d")
ddsCFM3 <- DESeq(ddsCFM,betaPrior=FALSE)
resCFM3 <- results(ddsCFM3)
resultsNames(ddsCFM3)
mcols(resCFM3, use.names=TRUE)
res3CFM <- results(ddsCFM3,"time_3.5d_vs_3d")
```

```
res3CFMna <- results(ddsCFM3,"time_3.5d_vs_3d")
res3CFM <-na.omit(res3CFM)
sum(res3CFM$padj < 0.05)
deg3CFM <- res3CFM[res3CFM$padj < 0.05, ]
deg3CFM$logFC.abs=abs(deg3CFM$log2FoldChange)
fc2_3 = deg3CFM[which(deg3CFM$logFC.abs>1),]
dim(fc2_3)
sum(fc2_3$log2FoldChange<0)
sum(fc2_3$log2FoldChange>0)
#write.csv(deg3CFM,file="DEGs3-3.5d")
plot(res3CFM$log2FoldChange, -log10(res3CFM$padj),pch=".",main="Volcano Plot for
3-3.5DAI Samples",col=ifelse(res3CFM$padj<0.05, "red","black"),xlab="log2(Fold
Change)",ylab="-log10(padj)")
sum(deg3CFM$log2FoldChange<0)
sum(deg3CFM$log2FoldChange>0)
###3.5-4.5###
colData(ddsCFM)$time <- relevel(colData(ddsCFM)$time, "3.5d")
ddsCFM3.5 <- DESeq(ddsCFM,betaPrior=FALSE)
resCFM3.5 <- results(ddsCFM3.5)
resultsNames(ddsCFM)
mcols(resCFM3.5, use.names=TRUE)
res3.5CFM <- results(ddsCFM3.5,"time_4.5d_vs_3.5d")
res3.5CFMna <- results(ddsCFM3.5,"time_4.5d_vs_3.5d")
res3.5CFM <-na.omit(res3.5CFM)
sum(res3.5CFM$padj < 0.05)
deg3.5CFM <- res3.5CFM[res3.5CFM$padj < 0.05, ]
deg3.5CFM$logFC.abs=abs(deg3.5CFM$log2FoldChange)
fc2_3.5 = deg3.5CFM[which(deg3.5CFM$logFC.abs>1),]
dim(fc2_3.5)
sum(fc2_3.5$log2FoldChange<0)
sum(fc2_3.5$log2FoldChange>0)
#write.csv(deg3.5CFM,file="DEGs4.5-3.5d")
#plot(res3.5CFM$log2FoldChange, -log10(res3.5CFM$padj),pch=".",main="Volcano
Plot for 3.5-4.5DAI Samples",col=ifelse(res3.5CFM$padj<0.05,
"red","black"),xlab="log2(Fold Change)",ylab="-log10(padj)")
sum(deg3.5CFM$log2FoldChange<0)
sum(deg3.5CFM$log2FoldChange>0)
###4.5-5.5###
colData(ddsCFM)$time <- relevel(colData(ddsCFM)$time, "4.5d")
ddsCFM4.5 <- DESeq(ddsCFM,betaPrior=FALSE)
resCFM4.5 <- results(ddsCFM4.5)
resultsNames(ddsCFM4.5)
mcols(resCFM4.5, use.names=TRUE)
res4.5CFM <- results(ddsCFM4.5,"time_5.5d_vs_4.5d")
res4.5CFMna <- results(ddsCFM4.5,"time_5.5d_vs_4.5d")
```

```
res4.5CFM <-na.omit(res4.5CFM)
sum(res4.5CFM$padj < 0.05)
deg4.5CFM <- res4.5CFM[res4.5CFM$padj < 0.05, ]
#write.csv(deg4.5CFM,file="DEGs5.5-4.5d")
deg4.5CFM$logFC.abs=abs(deg4.5CFM$log2FoldChange)
fc2_4.5 = deg4.5CFM[which(deg4.5CFM$logFC.abs>1),]
dim(fc2_4.5)
sum(fc2_4.5$log2FoldChange<0)
sum(fc2_4.5$log2FoldChange>0)
plot(res4.5CFM$log2FoldChange, -log10(res4.5CFM$padj),pch=".",main="Volcano Plot
for 4.5-5.5DAI Samples",col=ifelse(res4.5CFM$padj<0.05,
"red","black"),xlab="log2(Fold Change)",ylab="-log10(padj)")
sum(deg4.5CFM$log2FoldChange<0)
sum(deg4.5CFM$log2FoldChange>0)
##########################################################################
###################################
#    clustering
##########################################################################
###################################
###estimate theat using edgeR
library('edgeR')
library('mgcv')
#variance stabilising transformation
colData(ddsCFM)$time <- relevel(colData(ddsCFM)$time, "0d")
ddsCFM <- DESeq(ddsCFM,betaPrior=FALSE)
allcounts<-counts(ddsCFM, normalized=TRUE)

detach("package:DESeq",unload=TRUE)
vsdCFM <- varianceStabilizingTransformation(ddsCFM, blind=TRUE)

#get genes with highest variances
selectCFM <- order(-
rowVars(counts(ddsCFM,normalized=TRUE)),decreasing=TRUE)[1:3000]
selectCFM <-
order(rowMeans(counts(ddsCFM,normalized=TRUE)),decreasing=TRUE)[1:100]
#make a nice heatmap
colors <- colorpanel(75,"midnightblue","mediumseagreen","yellow")
heatmap.2(assay(vsdCFM)[selectCFM,], col=colors, dendrogram="both",
        scale="row", key=T, keysize=0.5, density.info="none",
        trace="none",cexCol=1.2, labRow=NA, RowSideColors=Label,
        lmat=rbind(c(5,0,4,0),c(3,1,2,0)), lhei=c(2.0,5.0),
        lwid=c(1.5,0.2,2.5,2.5))

heatmap.2(assay(vsdCFM)[selectCFM,], col = colors,dendrogram="both",
         scale="none", labRow=NA,Colv=NA,
```

```
        trace="none", margin=c(10, 6))

#make a heatmap of the expression patterns across time of the old differentially
expressed genes
oldGenes<-unique(read.table("OldgenesHitinNewExp.txt"))
rownames(oldGenes)<-oldGenes$V1

ddsoldGenesNewData = ddsCFM[rownames(ddsCFM)%in%rownames(oldGenes)]
vsdOld <- varianceStabilizingTransformation(ddsoldGenesNewData, blind=TRUE)
colors <- colorpanel(75,"midnightblue","mediumseagreen","yellow")
heatmap.2(assay(vsdOld), col = colors,dendrogram="both",
        scale="none", labRow=NA,
        trace="none", margin=c(10, 6))
countMeans<-matrix(0,dim(allcounts)[1],5)
countMeans[,1]<-rowMeans(allcounts[,1:2])
countMeans[,2]<-rowMeans(allcounts[,3:5])
countMeans[,3]<-rowMeans(allcounts[,6:8])
countMeans[,4]<-rowMeans(allcounts[,9:11])
countMeans[,5]<-rowMeans(allcounts[,12:14])
rownames(countMeans)<-rownames(allcounts)
colnames(countMeans)<-c("0dai","3dai","3.5dai","4.5dai","5.5dai")
use<-countMeans[rownames(countMeans)%in%rownames(oldGenes),]
#heatmap of log(average counts per condition)
heatmap.2(log(countMeans[rownames(countMeans)%in%rownames(oldGenes),]+1), col
= colors,dendrogram="both",
        scale="none", labRow=NA,
        trace="none", margin=c(10, 6))

##################Visualizations of the Data#########################


library('gplots')
cfm <- list("0-3"=rownames(deg0CFM),"3-3.5"=rownames(deg3CFM),"3.5-
4.5"=rownames(deg3.5CFM),"4.5-5.5"=rownames(deg4.5CFM))
venn(cfm)
inAll<- intersect(rownames(deg0CFM), rownames(deg3CFM))
inAll<- intersect(inAll, rownames(deg3.5CFM))
inAll<- intersect(inAll, rownames(deg4.5CFM))
inAllCFM<-subset(ddsCFM,rownames(ddsCFM) %in% inAll)


detach("package:DESeq", unload=TRUE)
rldinAllCFM <- rlogTransformation(inAllCFM, blind=TRUE)
vsdinAllCFM <- varianceStabilizingTransformation(inAllCFM, blind=TRUE)
```

```
heatmap.2(assay(vsdinAllCFM), col = colors,
        scale="none",labRow=NA,Rowv = FALSE, Colv = FALSE,dendrogram="none",
         trace="none", margin=c(10, 6))

rldCFM <- rlogTransformation(ddsCFM, blind=TRUE)
vsdCFM <- varianceStabilizingTransformation(ddsCFM, blind=TRUE)


#heatmaps of the count table (showing the 30 most highly expressed genes)
library("RColorBrewer")
library("gplots")
selectCFM <-
order(rowMeans(counts(ddsCFM,normalized=TRUE)),decreasing=TRUE)[1:30]
hmcolCFM <- colorRampPalette(brewer.pal(9, "GnBu"))(100)
#for raw counts
heatmap.2(counts(ddsCFM,normalized=TRUE)[selectCFM,], col = hmcol,
        Rowv = FALSE, Colv = FALSE, scale="none",
        dendrogram="none", trace="none", margin=c(10,6))
#for regularized log transformed data
heatmap.2(assay(rldCFM)[selectCFM,], col = hmcol,
        Rowv = FALSE, Colv = FALSE, scale="none",
        dendrogram="none", trace="none", margin=c(10, 6))
#for variance stabilizing transformed data
heatmap.2(assay(vsdCFM)[selectCFM,], col = hmcol,
        Rowv = FALSE, Colv = FALSE, scale="none",
        dendrogram="none", trace="none", margin=c(10, 6))
samplesCFM_bioRep<-data.frame(file_namesCFM,time,bioRep)
dataCFM_bioRep<-newCountDataSet(keep3,samplesCFM_bioRep)
countDataCFM_bioRep<-counts(dataCFM_bioRep)
colDataCFM_bioRep<-pData(dataCFM_bioRep)[,"time"]
ddsCFM_bioRep <- DESeqDataSetFromMatrix(countData = countDataCFM_bioRep,
                          colData = samplesCFM_bioRep,
                          design = ~ time+bioRep)
colData(ddsCFM_bioRep)$time <- factor(colData(ddsCFM_bioRep)$time,
                        levels=c("0dai","3dai","3.5dai","4.5dai","5.5dai"))
colData(ddsCFM_bioRep)$bioRep <- factor(colData(ddsCFM_bioRep)$bioRep,
                          levels=c("1","2","5"))
colData(ddsCFM_bioRep)$time <- relevel(colData(ddsCFM_bioRep)$time, "3dai")
colData(ddsCFM_bioRep)$bioRep <- relevel(colData(ddsCFM_bioRep)$bioRep, "1")
ddsCFM_bioRep <- DESeq(ddsCFM_bioRep,betaPrior=FALSE)
vsdCFM_bioRep <- varianceStabilizingTransformation(ddsCFM_bioRep, blind=TRUE)
rv = rowVars(assay(vsdCFM_bioRep))
select = order(rv, decreasing = TRUE)[seq_len(min(500, length(rv)))]
##get counts for genes in GA pathway and graph
GAgenes<-read.table('GAgeneHitsNewData.txt',header=FALSE)
```

```
GAcounts<-countMeans[rownames(countMeans)%in%GAgenes$V1,]
write.csv(GAcounts,file="GAgeneCountsBigE.csv")
dfga<-data.frame(GAcounts)
test<-read.csv("GAgraphs.csv",header=TRUE)
library('ggplot2')
c25 <- c("dodgerblue2","#E31A1C", # red
       "green4",
       "#6A3D9A", # purple
       "#FF7F00", # orange
       "black","gold1",
       "skyblue2","#FB9A99", # lt pink
       "palegreen2",
       "#CAB2D6", # lt purple
       "#FDBF6F", # lt orange
       "gray70", "khaki2",
       "maroon","orchid1","deeppink1","blue1","steelblue4",
       "darkturquoise","green1","yellow4","yellow3",
       "darkorange4","brown")
library('scales')
p<-ggplot(data=test, aes(x=time, y=counts,
group=geneName,color=geneName,shape=is.de,col =
factor(1:12)))+scale_x_continuous( breaks=c(0,3,3.5,4.5,5.5))+scale_y_continuous( brea
ks=c(0,1000,2000,3000,4000,5000)) +scale_colour_manual(values = c25)+
geom_line(size=1.0) +geom_point(size=3) + xlab("Days After Inoculation") +
ylab("Average Normalized Counts")+theme(axis.title.x =
element_text(face="bold",size=20))+theme(axis.title.y =
element_text(face="bold",size=20))
p+ theme_bw()
###do a Q-mode PCA (focuses on covariances and correlations between samples)###
pca = prcomp(t(assay(vsdCFM_bioRep)[select, ]))
summary(pca)
data = as.data.frame(pca$x)
ggplot(data, aes(PC1, PC2, color=time, shape=bioRep)) + geom_point(size=4) +
xlab("PC1: 83% variance") + ylab("PC2: 13%
variance")+scale_color_brewer(palette="Set1")+theme(axis.title.x =
element_text(face="bold",size=20))+theme(axis.title.y =
element_text(face="bold",size=20))
#######scatterplot matrix with red showing DEGs############
library('GGally')
distsVSDCFM <- dist(t(assay(vsdCFM)))
matCFM <- as.matrix(distsVSDCFM)
countsmat<-as.matrix(counts(ddsCFM),normalized=TRUE)
meanMat<-matrix(0,dim(counts(ddsCFM)),5)
meanMat[,1]<-rowMeans(countsmat[,1:2])
meanMat[,2]<-rowMeans(countsmat[,3:5])
```

```
meanMat[,3]<-rowMeans(countsmat[,6:8])
meanMat[,4]<-rowMeans(countsmat[,9:11])
meanMat[,5]<-rowMeans(countsmat[,12:14])
meanMat<-log2(meanMat)
meandf<-data.frame(meanMat)
meandf<-do.call(data.frame,lapply(meandf, function(x) replace(x, is.infinite(x),NA)))
meandf <-na.omit(meandf)
rownames(meanMat)<-rownames(countsmat)
colnames(meanMat)<-c("0DAI","3DAI","3.5DAI","4.5DAI","5.5DAI")
ggpairs(meandf)

data <- as.data.frame(meandf)
plot(log2(res$baseMeanA),log2(res$baseMeanB), pch=".", cex=.3, ylab="log2(baseMean)
+ACE", xlab="log2(baseMean) -ACE", col=ifelse(res$padj<0.01, "red","black"))

plotMatrix <- list(data = data, columns = columns, plots = ggpairsPlots,
                title = title, verbose = verbose, printInfo = printInfo,
                axisLabels = axisLabels)

rownames(matCFM) <- colnames(matCFM) <- with(colData(ddsCFM),
                                    paste(time, sep=" : "))
heatmap.2(matCFM, trace="none", col = rev(hmcol), margin=c(13, 13))
library('ggplot2')

#ggplot(geom_histogram(mapping = NULL, data = NULL, stat = "bin", position =
"stack", ...)
library('ggplot2')
library('RColorBrewer')
sort.deg0CFM <- fc2_0[order(fc2_0$log2FoldChange) , ]
sort.deg3CFM <- fc2_3[order(fc2_3$log2FoldChange) , ]
sort.deg3.5CFM <- fc2_3.5[order(fc2_3.5$log2FoldChange) , ]
sort.deg4.5CFM <- fc2_4.5[order(fc2_4.5$log2FoldChange) , ]
(sum(fc2_0$log2FoldChange>0)/(sum(fc2_0$log2FoldChange>0)+sum(fc2_0$log2Fold
Change<0)))*100
(sum(fc2_3$log2FoldChange>0)/(sum(fc2_3$log2FoldChange>0)+sum(fc2_3$log2Fold
Change<0)))*100
(sum(fc2_3.5$log2FoldChange>0)/(sum(fc2_3.5$log2FoldChange>0)+sum(fc2_3.5$log2
FoldChange<0)))*100
(sum(fc2_4.5$log2FoldChange>0)/(sum(fc2_4.5$log2FoldChange>0)+sum(fc2_4.5$log2
FoldChange<0)))*100
dim(fc2_0)
dim(fc2_3)
dim(fc2_3.5)
dim(fc2_4.5)
data<-matrix(0,(dim(fc2_0)+dim(fc2_3)+dim(fc2_3.5)+dim(fc2_4.5)),4)
```

```
data[,1]<-as.character(rep(c("0-3","3-3.5","3.5-4.5","4.5-
5.5"),c(length(fc2_0$log2FoldChange),length(fc2_3$log2FoldChange),length(fc2_3.5$lo
g2FoldChange),length(fc2_4.5$log2FoldChange))))
data[,3]<-
c(sort.deg0CFM$log2FoldChange,sort.deg3CFM$log2FoldChange,sort.deg3.5CFM$log
2FoldChange,sort.deg4.5CFM$log2FoldChange)
data[,2]<-as.numeric(c(seq(1, 13435, 1),seq(1, 2253, 1),seq(1, 4441, 1),seq(1, 4175, 1)))
data[,4]<-as.character(rep(c("51% up","98% up","80% up","75%
up"),c(length(fc2_0$log2FoldChange),length(fc2_3$log2FoldChange),length(fc2_3.5$lo
g2FoldChange),length(fc2_4.5$log2FoldChange))))
colnames(data)<-c("day","genes","log2fc","percent")
data<-data.frame(data)
data$genes<-as.numeric(as.character(data$genes))
data$log2fc<-as.numeric(as.character(data$log2fc))
hmcolCFM <- colorRampPalette(c("blue","red"))(100)
g<-ggplot(data, aes(x = day, y = genes, fill=log2fc)) +geom_tile()
+theme(legend.position = "top")+scale_fill_gradientn(colours = hmcolCFM,name="Log
base 2 Fold Change")+xlab("Time Interval (Days)")+ylab("Number of Genes")
g+annotate("text", x = 1, y = 14436, label = "51% ↑",size=10)+annotate("text", x = 2, y =
3254, label = "98% ↑",size=10)+annotate("text", x = 3, y = 5442, label = "80%
↑",size=10)+annotate("text", x = 4, y = 5176, label = "75% ↑",size=10)
#make fill like a heatmap
ggplot(test,aes(x=day))+geom_bar()+ylab("Genes") +theme(legend.position =
"top")+xlab("Time Interval (Days)")
print(plotPCA(rldCFM, intgroup=c("bioRep", "time")))
print(plotPCA(vsdCFM, intgroup=c( "time")))

plotDispEsts(vsdCFM,main="Dispersion Plot")

ressig = res0FM[res0FM$padj < 0.01,]
twenty<-subset(deg0FM,deg0FM$logFC.abs>4.32)
ten<-subset(deg0FM,deg0FM$logFC.abs>3.32)
four<-subset(deg0FM,deg0FM$logFC.abs>2)
two<-subset(deg0FM,deg0FM$logFC.abs>1)
plot(res0FM$baseMean,res0FM$log2FoldChange,log="x", pch=".",
cex=.3,ylab="log2(Fold Change)",xlab="baseMean",ylim=c(-15,15))
points(two$baseMean,two$log2FoldChange,pch='.',cex=3,ylim=c(-10,10),col='green')
points(four$baseMean,four$log2FoldChange,pch='.',cex=3,ylim=c(-10,10),col='orange')
points(ten$baseMean,ten$log2FoldChange,pch='.',cex=3,ylim=c(-10,10),col='blue')
points(twenty$baseMean,twenty$log2FoldChange,pch='.',cex=3,ylim=c(-10,10),col='red')
##########################make a matrix of 0,1,-1 & count
trends###################################
FMtrend<-matrix(0,dim(countsFM)[1],4)
rownames(FMtrend)<-rownames(countsFM)
```

```
colnames(FMtrend)<-c('0dai-3dai','3dai-3.5dai(-ACE)', '3.5dai(-ACE)-4.5dai(-
ACE)','4.5dai(-ACE)-5.5dai(-ACE)')
head(res3FM)
head(res3CFM)
#make a matrix of pvalues
CFMpval<-matrix(0,dim(keep3)[1],4)
rownames(CFMpval)<-rownames(keep3)
colnames(CFMpval)<-c('0dai-3dai','3dai-3.5dai(+ACE)', '3.5dai(+ACE)-
4.5dai(+ACE)','4.5dai(+ACE)-5.5dai(+ACE)')
CFMpval[,1]<-res0CFMna$padj
CFMpval[,2]<-res3CFMna$padj
CFMpval[,3]<-res3.5CFMna$padj
CFMpval[,4]<-res4.5CFMna$padj
head(CFMpval)

#make a matrix of log2FoldChange

CFMlog2fc<-matrix(0,dim(keep3)[1],4)
rownames(CFMlog2fc)<-rownames(keep3)
colnames(CFMlog2fc)<-c('0dai-3dai','3dai-3.5dai(+ACE)', '3.5dai(+ACE)-
4.5dai(+ACE)','4.5dai(+ACE)-5.5dai(+ACE)')
CFMlog2fc[,1]<-res0CFMna$log2FoldChange
CFMlog2fc[,2]<-res3CFMna$log2FoldChange
CFMlog2fc[,3]<-res3.5CFMna$log2FoldChange
CFMlog2fc[,4]<-res4.5CFMna$log2FoldChange
head(CFMlog2fc)
#make a matrix of test statistics

CFMstat<-matrix(0,dim(keep3)[1],4)
rownames(CFMstat)<-rownames(keep3)
colnames(CFMstat)<-c('0dai-3dai','3dai-3.5dai(+ACE)', '3.5dai(+ACE)-
4.5dai(+ACE)','4.5dai(+ACE)-5.5dai(+ACE)')
CFMstat[,1]<-res0CFMna$stat
CFMstat[,2]<-res3CFMna$stat
CFMstat[,3]<-res3.5CFMna$stat
CFMstat[,4]<-res4.5CFMna$stat
head(CFMstat)

#omit NAs
pval_2<-na.omit(CFMpval)
#keep only rows in CFMlog2fc which are in the pval_2 matrix also
CFMlog2fc_2<-subset(CFMlog2fc, rownames(CFMlog2fc) %in% rownames(pval_2))

CFMtrend<-matrix(0,dim(pval_2)[1],4)
rownames(CFMtrend)<-rownames(pval_2)
```

```
colnames(CFMtrend)<-c('0dai-3dai','3dai-3.5dai(+ACE)', '3.5dai(+ACE)-
4.5dai(+ACE)','4.5dai(+ACE)-5.5dai(+ACE)')

#loop through each gene in matrix
for(i in 1:length(CFMlog2fc_2[,1])){
  for(j in 1:4){
    if(CFMlog2fc_2[i,j]<=-1 && pval_2[i,j] < 0.05){ CFMtrend[i,j]<- -1 }
    if(pval_2[i,j] >= 0.05 ){ CFMtrend[i,j]<- 0 }
    if(CFMlog2fc_2[i,j]>-1 && CFMlog2fc_2[i,j]<1){ CFMtrend[i,j]<- 0 }
    if(CFMlog2fc_2[i,j]>=1 && pval_2[i,j] < 0.05){ CFMtrend[i,j]<- 1 }
  }
}
#write.csv(CFMtrend,file="geneTrendsCFMbigE")
###### make a matrix of possibilities###########
#for FM
grid<-expand.grid(c(-1,0,1),c(-1,0,1),c(-1,0,1),c(-1,0,1))
possibilities<-matrix(0,81,5)
for(i in 1:4){
  possibilities[,i]<-grid[,i]
}
colnames(possibilities)<-c('0dai-3dai','3dai-3.5dai', '3.5dai-4.5dai','4.5dai-5.5dai','total')

#For CFM
gridCFM<-expand.grid(c(-1,0,1),c(-1,0,1),c(-1,0,1),c(-1,0,1))
possibilitiesCFM<-matrix(0,81,5)
for(i in 1:4){
  possibilitiesCFM[,i]<-gridCFM[,i]
}
colnames(possibilitiesCFM)<-c('0dai-3dai','3dai-3.5dai', '3.5dai-4.5dai','4.5dai-
5.5dai','total')

for (i in 1:length(CFMtrend[,1])){
  tempvec1=as.vector(CFMtrend[i,1:4])
  for (j in 1:81){
    tempvec2<-as.vector(possibilitiesCFM[j,1:4])
    if (isTRUE(all.equal(tempvec1,tempvec2))){
      temp1=possibilitiesCFM[j,5]
      possibilitiesCFM[j,5]<-temp1+1
      break()
    }
  }
}
forGraphCFM<-matrix(0,81,6)
for(i in 1:81){
  forGraphCFM[i,2]<- possibilitiesCFM[i,1]
```

```
  temp1<-possibilitiesCFM[i,1]
  temp2<-possibilitiesCFM[i,2]
  temp3<-possibilitiesCFM[i,3]
  temp4<-possibilitiesCFM[i,4]
  forGraphCFM[i,3]<-(forGraphCFM[i,2])+temp2
  forGraphCFM[i,4]<-(forGraphCFM[i,3])+temp3
  forGraphCFM[i,5]<-(forGraphCFM[i,4])+temp4
  forGraphCFM[i,6]<- possibilitiesCFM[i,5]
}
colnames(forGraphCFM)<-c('0dai','3dai','3.5dai', '4.5dai','5.5dai','total')
write.csv(forGraphCFM, file="patternsCFM1.csv")
######plots#######
#-ACE##
library("ggplot2")
library('graphics')
library("reshape")


df<-read.table("patternsCFM1.txt",header=TRUE)
library( RColorBrewer)
df1=df[df$ofInterest!=0,]
df2<-data.frame(df1[1:8,1:7])
colnames(df2)<-c("pattern","0","3","3.5","4.5","5.5","total")
c25 <- c("dodgerblue2","#E31A1C", # red
        "green4",
        "#6A3D9A", # purple
        "#FF7F00", # orange
        "black","gold1",
        "skyblue2","#FB9A99", # lt pink
        "palegreen2",
        "#CAB2D6", # lt purple
        "#FDBF6F", # lt orange
        "gray70", "khaki2",
        "maroon","orchid1","deeppink1","blue1","steelblue4",
        "darkturquoise","green1","yellow4","yellow3",
        "darkorange4","brown")
dfm<-melt(df2,id.vars=c("total","pattern"))
dfm2 <- dfm
dfm2$pattern <- factor(dfm2$pattern)
p<-ggplot(dfm2, aes(x=variable, y=value, colour=factor(total),group=pattern),)
+theme(panel.grid.minor=element_blank(), panel.grid.major=element_blank())
p<-p+scale_y_continuous( breaks=c(-4,-3,-2,-1,0,1,2,3,4))+  xlab("Time(Days After
Innoculation)")
p<- p+  ylab("Gene Expression Pattern") + ggtitle("Gene Expression Patterns Across
Time (+ACE)")
```

```
p<- p+scale_colour_manual(values =
c25)+geom_line( aes(linetype=factor(total)),size=1.4)

p

plot(df, ylab="Gene Expression Pattern",main="Gene Expression Patterns Across
Time",axes=FALSE,type="l",sub="-ACE",
    xlab="Time(Days After Innoculation)"     )
axis(1, at=1:5, lab=c("0dai","3dai","3.5dai","4.5dai","5.5dai"))
axis(2, las=1, at=c(-4,-3,-2,-1,0,1,2,3,4))
box()

#############################################################################

detach("package:DESeq", unload=TRUE)

pvalCFM<-matrix(0,dim(res0CFMna)[1],4)
pvalCFM[,1]<-res0CFMna$padj
pvalCFM[,2]<-res3CFMna$padj
pvalCFM[,3]<-res3.5CFMna$padj
pvalCFM[,4]<-res4.5CFMna$padj
colnames(pvalCFM)<-c('0DAI','3DAI','4.5DAI','5.5DAI')
rownames(pvalCFM)<-rownames(res0CFMna)
head(pvalCFM)
library("vsn")
par(mfrow=c(1,3))
notAllZeroCFM <- (rowSums(counts(ddsCFM))>0)
meanSdPlot(log2(counts(ddsCFM,normalized=TRUE)[notAllZeroCFM,] + 1),ylim =
c(0,2.5))
meanSdPlot(assay(rldCFM[notAllZeroCFM,]), ylim = c(0,2.5))
meanSdPlot(assay(vsdCFM[notAllZeroCFM,]), ylim = c(0,2.5))

#make GO barchart
setwd("~/Desktop")
library( ggplot2 )

gos<-read.csv("ForGOgraph.csv", header=TRUE)

p<-ggplot(data=gos, aes(x=factor(GO),y=Percent,fill=factor(GO)))
+geom_bar(stat="identity")
p<-p + scale_y_continuous( breaks=c(5,10,15,20,25,30,35,40)) + theme_bw()
c20 <- c("dodgerblue2","#E31A1C", # red
        "green4",
        "#6A3D9A", # purple
```

```
        "#FF7F00", # orange
        "gold1",
        "skyblue2","#FB9A99", # lt pink
        "palegreen2",
        "#CAB2D6", # lt purple
        "#FDBF6F", # lt orange

        "maroon","orchid1","deeppink1","blue1","steelblue4",
        "darkturquoise","green1","yellow4","yellow3",
        "darkorange4")
c15 <- c("#CCFFFF","#660000",
        "#003300",
        "#6633FF", # purple
        "#990033", # orange
        "#00CC99","#FF0066", # lt pink
        "#99FF66",
        "#660066", # lt purple
        "black","#CCCCCC","#CC9900","#006699","#336666",
        "#003333")
p<- p+scale_fill_manual(values = c15)
p

#other GO plot for genes expressed at 0 days
go_zeroH<-read.csv("0dayGOslimUseinGraph.csv", header=TRUE)
p0 <- ggplot(data=go_zeroH,
aes(x=factor(Term.Name.),y=X.Seq,fill=factor(Term.Name.)))
p0 <- p0+geom_bar(stat="identity",show_guide = FALSE)
p0 <- p0 + theme(axis.text.x  = element_text(hjust=1, vjust=0.3, angle=90,
colour='black'),axis.text.y  = element_text(colour='black'))
p0 <- p0 + ylab("Number of Sequences") + xlab("Biological Process GO Term")
p0 <- p0 +geom_text(aes(y=X.Seq, ymax=X.Seq, label=X.Seq),position=
position_dodge(width=0.9), vjust=-.5, size=3)
p0

#make GA graphs
library( ggplot2 )

model<-read.csv("modelGenes_GAgraphs.csv",header=TRUE)
p<-ggplot(data=model, aes(x=time, y=counts,
group=geneName,color=geneName,shape=is.de,col =
factor(1:12)))+scale_x_continuous( breaks=c(0,3,3.5,4.5,5.5))+scale_y_continuous( brea
ks=c(0,1000,2000,3000,4000,5000)) +scale_colour_manual(values = c25)+
geom_line(size=1.0) +geom_point(size=3) + xlab("Days After Inoculation") +
ylab("Average Normalized Counts")+theme(axis.title.x =
```

```
element_text(face="bold",size=20))+theme(axis.title.y =
element_text(face="bold",size=20))
limits <- aes(ymax = counts + stdev, ymin=counts - stdev)
p + geom_errorbar(limits, width=0.25)

tfs<-read.csv("GA-related_transcriptionfactors_forgraphs.csv",header=TRUE)
p<-ggplot(data=tfs, aes(x=time, y=counts,
group=geneName,color=geneName,shape=is.de,col =
factor(1:12)))+scale_x_continuous( breaks=c(0,3,3.5,4.5,5.5))+scale_y_continuous( brea
ks=c(0,1000,2000,3000,4000,5000)) +scale_colour_manual(values = c25)+
geom_line(size=1.0) +geom_point(size=3) + xlab("Days After Inoculation") +
ylab("Average Normalized Counts")+theme(axis.title.x =
element_text(face="bold",size=20))+theme(axis.title.y =
element_text(face="bold",size=20))
limits <- aes(ymax = counts + stdev, ymin=counts - stdev)
p + geom_errorbar(limits, width=0.25)

sigtrans<-read.csv("signalTrans_forgraph.csv",header=TRUE)
p<-ggplot(data=sigtrans, aes(x=time, y=counts,
group=geneName,color=geneName,shape=is.de,col =
factor(1:12)))+scale_x_continuous( breaks=c(0,3,3.5,4.5,5.5))+scale_y_continuous( brea
ks=c(0,1000,2000,3000,4000,5000)) +scale_colour_manual(values = c25)+
geom_line(size=1.0) +geom_point(size=3) + xlab("Days After Inoculation") +
ylab("Average Normalized Counts")+theme(axis.title.x =
element_text(face="bold",size=20))+theme(axis.title.y =
element_text(face="bold",size=20))
limits <- aes(ymax = counts + stdev, ymin=counts - stdev)
p + geom_errorbar(limits, width=0.25)

biosyn<-read.csv("biosyn_forgraphs.csv",header=TRUE)
p<-ggplot(data=biosyn, aes(x=time, y=counts,
group=geneName,color=geneName,shape=is.de,col =
factor(1:12)))+scale_x_continuous( breaks=c(0,3,3.5,4.5,5.5))+scale_y_continuous( brea
ks=c(0,1000,2000,3000,4000,5000)) +scale_colour_manual(values = c25)+
geom_line(size=1.0) +geom_point(size=3) + xlab("Days After Inoculation") +
ylab("Average Normalized Counts")+theme(axis.title.x =
element_text(face="bold",size=20))+theme(axis.title.y =
element_text(face="bold",size=20))
limits <- aes(ymax = counts + stdev, ymin=counts - stdev)
p + geom_errorbar(limits, width=0.25)


#GO enrichment test with GoSeq
setwd("~/Desktop")
####GO enrichment test##################
```

```
source("http://bioconductor.org/biocLite.R")
library(goseq)
library(GO.db)
library("biomaRt")

#median of isoform lengths
lengthData<-read.table("allNames_medianLen",row.names=1)

#go annotation using blast results against blastx
#format: comp10000<TAB>GO:1919191, one comp-go pair a line
go <- read.table("GOtermsBigE.txt", header=FALSE, sep="\t", fill=TRUE)
head(go)
#get GOslim terms from BioMart
ensembl <- useMart("plants_mart_24",dataset="athaliana_eg_gene")
slim = useMart("ensembl",dataset="hsapiens_gene_ensembl")
go_slim<-getBM(attributes="goslim_goa_accession",mart=slim)[,1]
#go_slim<-read.csv("go_slim.csv",header=TRUE)
#go_slim<-as.vector(go_slim[,2])

#filter GO terms to keep only GOslim terms
go_slim2cat<-subset(go, go[,2] %in% go_slim)
#names of all comp names kept in DEG analysis
keep <- read.table('BigEallnames.txt')

#all DEGs identified
genes0_3<-read.table("0-3daynames.txt")
genes3_3.5<-read.table("3-3.5daynames.txt")
genes3.5_4.5<-read.table("3.5-4.5daynames.txt")
genes4.5_5.5<-read.table("4.5-5.5daynames.txt")
genesLRT<-read.table("LRTnames.txt")
genesgreater0.3CPM<-read.table("greater0.3CPMnames.txt")
genesless0.3CPM<-read.table("less0.3CPMnames.txt")
pattern71G<-read.table("71Gpatternnames.txt")
pattern504G<-read.table("504Gpatternnames.txt")
pattern570G<-read.table("570Gpatternnames.txt")
pattern834G<-read.table("834Gpatternnames.txt")
pattern4738G<-read.table("4738Gpatternnames.txt")
pattern4806G<-read.table("4806Gpatternnames.txt")
pattern9981G<-read.table("9981Gpatternnames.txt")

new71<-go[go[,1]%in%pattern71G[,1],]
new504<-go[go[,1]%in%pattern504G[,1],]
new570<-go[go[,1]%in%pattern570G[,1],]
new834<-go[go[,1]%in%pattern834G[,1],]
new4738<-go[go[,1]%in%pattern4738G[,1],]
```

```
new4806<-go[go[,1]%in%pattern4806G[,1],]
new9981<-go[go[,1]%in%pattern9981G[,1],]

#write.table(new71,file="Pattern71GO.txt")
#write.table(new504,file="Pattern504GO.txt")
#write.table(new570,file="Pattern570GO.txt")
#write.table(new834,file="Pattern834GO.txt")
#write.table(new4738,file="Pattern4738GO.txt")
#write.table(new4806,file="Pattern4806GO.txt")
#write.table(new9981,file="Pattern9981GO.txt")

pattern71=as.integer(keep[,1]%in%pattern71G[,1])
names(pattern71)=keep[,1]
head(pattern71)
pattern504=as.integer(keep[,1]%in%pattern504G[,1])
names(pattern504)=keep[,1]
head(pattern504)
pattern570=as.integer(keep[,1]%in%pattern570G[,1])
names(pattern570)=keep[,1]
head(pattern570)
pattern834=as.integer(keep[,1]%in%pattern834G[,1])
names(pattern834)=keep[,1]
head(pattern834)
pattern4738=as.integer(keep[,1]%in%pattern4738G[,1])
names(pattern4738)=keep[,1]
head(pattern4738)
pattern4806=as.integer(keep[,1]%in%pattern4806G[,1])
names(pattern4806)=keep[,1]
head(pattern4806)
pattern9981=as.integer(keep[,1]%in%pattern9981G[,1])
names(pattern9981)=keep[,1]
head(pattern9981)
genes_great=as.integer(keep[,1]%in%genesgreater0.3CPM[,1])
names(genes_great)=keep[,1]
head(genes_great)
genes_less=as.integer(keep[,1]%in%genesless0.3CPM[,1])
names(genes_less)=keep[,1]
head(genes_less)
genes_LRT=as.integer(keep[,1]%in%genesLRT[,1])
names(genes_LRT)=keep[,1]
head(genes_LRT)
genes0=as.integer(keep[,1]%in%genes0_3[,1])
names(genes0)=keep[,1]
head(genes0)
genes3=as.integer(keep[,1]%in%genes3_3.5[,1])
```

```
names(genes3)=keep[,1]
head(genes3)
genes3.5=as.integer(keep[,1]%in%genes3.5_4.5[,1])
names(genes3.5)=keep[,1]
head(genes3.5)
genes4.5=as.integer(keep[,1]%in%genes4.5_5.5[,1])
names(genes4.5)=keep[,1]
head(genes4.5)

bias_4.5=lengthData[rownames(lengthData)%in%names(genes4.5),]
names(bias_4.5) = rownames(lengthData)[rownames(lengthData)%in%names(genes4.5)]
head(bias_4.5)
bias_0=lengthData[rownames(lengthData)%in%names(genes0),]
names(bias_0) = rownames(lengthData)[rownames(lengthData)%in%names(genes0)]
head(bias_0)
bias_3=lengthData[rownames(lengthData)%in%names(genes3),]
names(bias_3) = rownames(lengthData)[rownames(lengthData)%in%names(genes3)]
head(bias_3)
bias_3.5=lengthData[rownames(lengthData)%in%names(genes3.5),]
names(bias_3.5) = rownames(lengthData)[rownames(lengthData)%in%names(genes3.5)]
head(bias_3.5)
bias_71=lengthData[rownames(lengthData)%in%names(pattern71),]
names(bias_71) = rownames(lengthData)[rownames(lengthData)%in%names(pattern71)]
head(bias_71)
bias_504=lengthData[rownames(lengthData)%in%names(pattern504),]
names(bias_504) =
rownames(lengthData)[rownames(lengthData)%in%names(pattern504)]
head(bias_504)
bias_570=lengthData[rownames(lengthData)%in%names(pattern570),]
names(bias_570) =
rownames(lengthData)[rownames(lengthData)%in%names(pattern570)]
head(bias_570)
bias_834=lengthData[rownames(lengthData)%in%names(pattern834),]
names(bias_834) =
rownames(lengthData)[rownames(lengthData)%in%names(pattern834)]
head(bias_834)
bias_4738=lengthData[rownames(lengthData)%in%names(pattern4738),]
names(bias_4738) =
rownames(lengthData)[rownames(lengthData)%in%names(pattern4738)]
head(bias_4738)
bias_4806=lengthData[rownames(lengthData)%in%names(pattern4806),]
names(bias_4806) =
rownames(lengthData)[rownames(lengthData)%in%names(pattern4806)]
head(bias_4806)
bias_9981=lengthData[rownames(lengthData)%in%names(pattern9981),]
```

```
names(bias_9981) =
rownames(lengthData)[rownames(lengthData)%in%names(pattern9981)]
head(bias_9981)
bias_less=lengthData[rownames(lengthData)%in%names(genes_less),]
names(bias_less) =
rownames(lengthData)[rownames(lengthData)%in%names(genes_less)]
head(bias_less)
bias_great=lengthData[rownames(lengthData)%in%names(genes_great),]
names(bias_great) =
rownames(lengthData)[rownames(lengthData)%in%names(genes_great)]
head(bias_great)
bias_LRT=lengthData[rownames(lengthData)%in%names(genes_LRT),]
names(bias_LRT) =
rownames(lengthData)[rownames(lengthData)%in%names(genes_LRT)]
head(bias_LRT)

pwf_p71 = nullp(pattern71,bias.data=bias_71)
pwf_p504 = nullp(pattern504,bias.data=bias_504)
pwf_p570 = nullp(pattern570,bias.data=bias_570)
pwf_p834 = nullp(pattern834,bias.data=bias_834)
pwf_p4738 = nullp(pattern4738,bias.data=bias_4738)
pwf_p4806 = nullp(pattern4806,bias.data=bias_4806)
pwf_p9981 = nullp(pattern9981,bias.data=bias_9981)
pwf_0 = nullp(genes0,bias.data=bias_0)
pwf_3 = nullp(genes3,bias.data=bias_3)
pwf_3.5 = nullp(genes3.5,bias.data=bias_3.5)
pwf_4.5 = nullp(genes4.5,bias.data=bias_4.5)
pwf_LRT = nullp(genes_LRT,bias.data=bias_LRT)
pwf_great = nullp(genes_great,bias.data=bias_great)
pwf_less = nullp(genes_less,bias.data=bias_less)

go <- read.table("GOtermsBigE.txt", header=FALSE, sep="\t", fill=TRUE)
GO.wall.p71 <- goseq(pwf_p71, gene2cat=go)
GO.wall.p71=goseq(pwf_p71,gene2cat=go_slim2cat)
enriched.GO.wall.p71 = GO.wall.p71$category[GO.wall.p71$over_represented_pvalue
<=0.05]
sink(file="enrichedGOannot_p71GO0.5.txt")
for(go in enriched.GO.wall.p71[1:length(enriched.GO.wall.p71)]){print(GOTERM[[go]])
                        cat("--------------------------------------\n")
}
sink()

GO.wall.M <- goseq(pwf_less, gene2cat=go
              GO.wall.H <- goseq(Hpwf, gene2cat=go)
              GO.wall.M=goseq(Mpwf,gene2cat=go_slim2cat)
```

```
GO.wall.H=goseq(Hpwf,gene2cat=go_slim2cat)
head(GO.wall.M)
enriched.GO.M =
GO.wall.M$category[GO.wall.M$over_represented_pvalue <=0.05]
enriched.GO.H = GO.wall.H$category[GO.wall.H$over_represented_pvalue
<=0.05]
head(enriched.GO.M)
#print in a file
sink(file="enrichedGOannot_lessGO0.5.txt")
for(go in
enriched.GO.wall.less[1:length(enriched.GO.wall.less)]){print(GOTERM[[go]])
 cat("--------------------------------------\n")
}
sink()
```

Appendix B Time-Course GO enrichment Results

Table B.1. Enrichment analysis results for time-course RNA-Seq experiment. The "Enriched in DEGs" column shows the pairs of time-points in which differentially expressed genes show an enrichment for the given GO term. BP=biological process, MF=molecular function, CC=cellular component

GO term	Enriched in DEGs	Ontology	GO term Description
GO:0006633	0-3, 3-3.5, 3.5-4.5, 4.5-5.5	BP	fatty acid biosynthetic process
GO:0016310	0-3, 3-3.5, 3.5-4.5, 4.5-5.5	BP	phosphorylation
GO:0055085	0-3, 3-3.5, 3.5-4.5, 4.5-5.5	BP	transmembrane transport
GO:0006468	0-3, 3-3.5, 3.5-4.5	BP	protein phosphorylation
GO:0042546	0-3, 3-3.5, 3.5-4.5	BP	cell wall biogenesis
GO:0003333	0-3	BP	amino acid transmembrane transport
GO:0006950	0-3	BP	response to stress
GO:0007000	0-3	BP	nucleolus organization
GO:0009082	0-3	BP	branched-chain amino acid biosynthetic process
GO:0010182	0-3	BP	sugar mediated signaling pathway
GO:0010206	0-3	BP	photosystem II repair
GO:0019538	0-3	BP	protein metabolic process
GO:0030244	0-3	BP	cellulose biosynthetic process
GO:0048829	0-3	BP	root cap development
GO:0080156	0-3	BP	mitochondrial mRNA modification
GO:0046148	3-3., 3.5-4.5, 4.5-5.5	BP	pigment biosynthetic process
GO:0005975	3-3.5, 3.5-4.5, 4.5-5.5	BP	carbohydrate metabolic process

Table B.1 Continued

GO:0006351	3-3.5, 3.5-4.5, 4.5-5.5	BP	transcription, DNA-dependent
GO:0006979	3-3.5, 3.5-4.5, 4.5-5.5	BP	response to oxidative stress
GO:0008152	3-3.5, 3.5-4.5, 4.5-5.5	BP	metabolic process
GO:0009765	3-3.5, 3.5-4.5, 4.5-5.5	BP	photosynthesis, light harvesting
GO:0009768	3-3.5, 3.5-4.5, 4.5-5.5	BP	photosynthesis, light harvesting in photosystem I
GO:0015979	3-3.5, 3.5-4.5, 4.5-5.5	BP	photosynthesis
GO:0016126	3-3.5, 3.5-4.5, 4.5-5.5	BP	sterol biosynthetic process
GO:0016132	3-3.5, 3.5-4.5, 4.5-5.5	BP	brassinosteroid biosynthetic process
GO:0042545	3-3.5, 3.5-4.5, 4.5-5.5	BP	cell wall modification
GO:0046274	3-3.5, 3.5-4.5, 4.5-5.5	BP	lignin catabolic process
GO:0055114	3-3.5, 3.5-4.5, 4.5-5.5	BP	oxidation-reduction process
GO:0000079	3-3.5, 3.5-4.5	BP	regulation of cyclin-dependent protein kinase activity
GO:0000226	3-3.5, 3.5-4.5	BP	microtubule cytoskeleton organization
GO:0000280	3-3.5, 3.5-4.5	BP	nuclear division
GO:0006084	3-3.5, 3.5-4.5	BP	acetyl-CoA metabolic process
GO:0006200	3-3.5, 3.5-4.5	BP	ATP catabolic process
GO:0006334	3-3.5, 3.5-4.5	BP	nucleosome assembly
GO:0006556	3-3.5, 3.5-4.5	BP	S-adenosylmethionine biosynthetic process
GO:0006559	3-3.5, 3.5-4.5	BP	L-phenylalanine catabolic process
GO:0006869	3-3.5, 3.5-4.5	BP	lipid transport
GO:0006952	3-3.5, 3.5-4.5	BP	defense response

Table B.1. Continuted

GO:0007049	3-3.5, 3.5-4.5	BP	cell cycle
GO:0007169	3-3.5, 3.5-4.5	BP	transmembrane receptor protein tyrosine kinase signaling pathway
GO:0009652	3-3.5, 3.5-4.5	BP	thigmotropism
GO:0009800	3-3.5, 3.5-4.5	BP	cinnamic acid biosynthetic process
GO:0009807	3-3.5, 3.5-4.5	BP	lignan biosynthetic process
GO:0010114	3-3.5, 3.5-4.5	BP	response to red light
GO:0010218	3-3.5, 3.5-4.5	BP	response to far red light
GO:0010583	3-3.5, 3.5-4.5	BP	response to cyclopentenone
GO:0016458	3-3.5, 3.5-4.5	BP	gene silencing
GO:0016572	3-3.5, 3.5-4.5	BP	histone phosphorylation
GO:0030001	3-3.5, 3.5-4.5	BP	metal ion transport
GO:0030865	3-3.5, 3.5-4.5	BP	cortical cytoskeleton organization
GO:0043086	3-3.5, 3.5-4.5	BP	negative regulation of catalytic activity
GO:0048281	3-3.5, 3.5-4.5	BP	inflorescence morphogenesis
GO:0048443	3-3.5, 3.5-4.5	BP	stamen development
GO:0048451	3-3.5, 3.5-4.5	BP	petal formation
GO:0048453	3-3.5, 3.5-4.5	BP	sepal formation
GO:0051225	3-3.5, 3.5-4.5	BP	spindle assembly
GO:0051301	3-3.5, 3.5-4.5	BP	cell division
GO:0090116	3-3.5, 3.5-4.5	BP	C-5 methylation of cytosine
GO:0000041	3-3.5, 3.5-4.5s,4.5-5.5	BP	transition metal ion transport
GO:0006073	3-3.5,3.5-4.5,4.5-5.5	BP	cellular glucan metabolic process
GO:0006006	3-3.5,3.5-4.5	BP	glucose metabolic process
GO:0007018	3-3.5,3.5-4.5	BP	microtubule-based movement
GO:0010389	3-3.5,3.5-4.5	BP	regulation of G2/M transition of mitotic cell cycle

Table B.1 Continued

GO:0000911	3-3.5,4.5-5.5	BP	cytokinesis by cell plate formation
GO:0005985	3-3.5	BP	sucrose metabolic process
GO:0006260	3-3.5	BP	DNA replication
GO:0006265	3-3.5	BP	DNA topological change
GO:0006270	3-3.5	BP	DNA-dependent DNA replication initiation
GO:0006275	3-3.5	BP	regulation of DNA replication
GO:0006306	3-3.5	BP	DNA methylation
GO:0006342	3-3.5	BP	chromatin silencing
GO:0006346	3-3.5	BP	methylation-dependent chromatin silencing
GO:0006820	3-3.5	BP	anion transport
GO:0006873	3-3.5	BP	cellular ion homeostasis
GO:0006882	3-3.5	BP	cellular zinc ion homeostasis
GO:0007067	3-3.5	BP	mitosis
GO:0008283	3-3.5	BP	cell proliferation
GO:0009186	3-3.5	BP	deoxyribonucleoside diphosphate metabolic process
GO:0009270	3-3.5	BP	response to humidity
GO:0009585	3-3.5	BP	red, far-red light phototransduction
GO:0009698	3-3.5	BP	phenylpropanoid metabolic process
GO:0009909	3-3.5	BP	regulation of flower development
GO:0010037	3-3.5	BP	response to carbon dioxide
GO:0010103	3-3.5	BP	stomatal complex morphogenesis
GO:0010119	3-3.5	BP	regulation of stomatal movement
GO:0010193	3-3.5	BP	response to ozone
GO:0010215	3-3.5	BP	cellulose microfibril organization

Table B.1 Continued

GO:0010223	3-3.5	BP	secondary shoot formation
GO:0010417	3-3.5	BP	glucuronoxylan biosynthetic process
GO:0010584	3-3.5	BP	pollen exine formation
GO:0030261	3-3.5	BP	chromosome condensation
GO:0031047	3-3.5	BP	gene silencing by RNA
GO:0031048	3-3.5	BP	chromatin silencing by small RNA
GO:0031669	3-3.5	BP	cellular response to nutrient levels
GO:0034219	3-3.5	BP	carbohydrate transmembrane transport
GO:0034968	3-3.5	BP	histone lysine methylation
GO:0048229	3-3.5	BP	gametophyte development
GO:0050891	3-3.5	BP	multicellular organismal water homeostasis
GO:0051567	3-3.5	BP	histone H3-K9 methylation
GO:0007389	3.5-4.5, 4.5-5.5	BP	pattern specification process
GO:0008361	3.5-4.5, 4.5-5.5	BP	regulation of cell size
GO:0009725	3.5-4.5, 4.5-5.5	BP	response to hormone stimulus
GO:0009926	3.5-4.5, 4.5-5.5	BP	auxin polar transport
GO:0009954	3.5-4.5, 4.5-5.5	BP	proximal/distal pattern formation
GO:0009969	3.5-4.5, 4.5-5.5	BP	xyloglucan biosynthetic process
GO:0010054	3.5-4.5, 4.5-5.5	BP	trichoblast differentiation
GO:0010075	3.5-4.5, 4.5-5.5	BP	regulation of meristem growth
GO:0015706	3.5-4.5, 4.5-5.5	BP	nitrate transport
GO:0018298	3.5-4.5, 4.5-5.5	BP	protein-chromophore linkage

Table B.1 Continued

GO:0043481	3.5-4.5, 4.5-5.5	BP	anthocyanin accumulation in tissues in response to UV light
GO:0000904	3.5-4.5,4.5-5.5	BP	cell morphogenesis involved in differentiation
GO:0009734	3.5-4.5,4.5-5.5	BP	auxin mediated signaling pathway
GO:0000271	3.5-4.5	BP	polysaccharide biosynthetic process
GO:0000302	3.5-4.5	BP	response to reactive oxygen species
GO:0006108	3.5-4.5	BP	malate metabolic process
GO:0006536	3.5-4.5	BP	glutamate metabolic process
GO:0006598	3.5-4.5	BP	polyamine catabolic process
GO:0006629	3.5-4.5	BP	lipid metabolic process
GO:0006817	3.5-4.5	BP	phosphate ion transport
GO:0006885	3.5-4.5	BP	regulation of pH
GO:0007017	3.5-4.5	BP	microtubule-based process
GO:0007165	3.5-4.5	BP	signal transduction
GO:0007623	3.5-4.5	BP	circadian rhythm
GO:0009637	3.5-4.5	BP	response to blue light
GO:0009664	3.5-4.5	BP	plant-type cell wall organization
GO:0009832	3.5-4.5	BP	plant-type cell wall biogenesis
GO:0009932	3.5-4.5	BP	cell tip growth
GO:0010103	3.5-4.5	BP	stomatal complex morphogenesis
GO:0010411	3.5-4.5	BP	xyloglucan metabolic process
GO:0010817	3.5-4.5	BP	regulation of hormone levels
GO:0015770	3.5-4.5	BP	sucrose transport
GO:0032774	3.5-4.5	BP	RNA biosynthetic process

Table B.1 Continued

GO:0043132	3.5-4.5	BP	NAD transport
GO:0044375	3.5-4.5	BP	regulation of peroxisome size
GO:0051258	3.5-4.5	BP	protein polymerization
GO:0070417	3.5-4.5	BP	cellular response to cold
GO:0071484	3.5-4.5	BP	cellular response to light intensity
GO:0000038	4.5-5.5	BP	very long-chain fatty acid metabolic process
GO:0006072	4.5-5.5	BP	glycerol-3-phosphate metabolic process
GO:0006090	4.5-5.5	BP	pyruvate metabolic process
GO:0006200	4.5-5.5	BP	ATP catabolic process
GO:0006278	4.5-5.5	BP	RNA-dependent DNA replication
GO:0006723	4.5-5.5	BP	cuticle hydrocarbon biosynthetic process
GO:0006810	4.5-5.5	BP	transport
GO:0006817	4.5-5.5	BP	phosphate ion transport
GO:0009944	4.5-5.5	BP	polarity specification of adaxial/abaxial axis
GO:0010025	4.5-5.5	BP	wax biosynthetic process
GO:0010315	4.5-5.5	BP	auxin efflux
GO:0015074	4.5-5.5	BP	DNA integration
GO:0015696	4.5-5.5	BP	ammonium transport
GO:0015995	4.5-5.5	BP	chlorophyll biosynthetic process
GO:0019684	4.5-5.5	BP	photosynthesis, light reaction
GO:0019752	4.5-5.5	BP	carboxylic acid metabolic process
GO:0019932	4.5-5.5	BP	second-messenger-mediated signaling
GO:0030418	4.5-5.5	BP	nicotianamine biosynthetic process
GO:0042128	4.5-5.5	BP	nitrate assimilation
GO:0042773	4.5-5.5	BP	ATP synthesis coupled electron transport

Tabe B.1 Continued

GO:0043447	4.5-5.5	BP	alkane biosynthetic process
GO:0046168	4.5-5.5	BP	glycerol-3-phosphate catabolic process
GO:0046482	4.5-5.5	BP	para-aminobenzoic acid metabolic process
GO:0048235	4.5-5.5	BP	pollen sperm cell differentiation
GO:0051188	4.5-5.5	BP	cofactor biosynthetic process
GO:0060964	4.5-5.5	BP	regulation of gene silencing by miRNA
GO:0072488	4.5-5.5	BP	ammonium transmembrane transport
GO:0090305	4.5-5.5	BP	nucleic acid phosphodiester bond hydrolysis
GO:0003989	0-3	MF	acetyl-CoA carboxylase activity
GO:0004712	0-3	MF	protein serine/threonine/tyrosine kinase activity
GO:0004748	0-3	MF	ribonucleoside-diphosphate reductase activity, thioredoxin
GO:0004806	0-3	MF	triglyceride lipase activity
GO:0004965	0-3	MF	G-protein coupled GABA receptor activity
GO:0005249	0-3	MF	voltage-gated potassium channel activity
GO:0019894	0-3	MF	kinesin binding
GO:0004185	0-3, 3-3.5, 3.5-4.5	MF	serine-type carboxypeptidase activity
GO:0004672	0-3, 3-3.5, 3.5-4.5	MF	protein kinase activity
GO:0004674	0-3, 3-3.5, 3.5-4.5	MF	protein serine/threonine kinase activity

Table B.1 Continued

GO:0016772	0-3, 3-3.5, 3.5-4.5	MF	transferase activity, transferring phosphorus-containing groups
GO:0005215	0-3, 3-3.5, 3.5-4.5, 4.5-5.5	MF	transporter activity
GO:0005506	0-3, 3-3.5, 3.5-4.5, 4.5-5.5	MF	iron ion binding
GO:0009055	0-3, 3-3.5, 3.5-4.5, 4.5-5.5	MF	electron carrier activity
GO:0016491	0-3, 3-3.5, 3.5-4.5, 4.5-5.5	MF	oxidoreductase activity
GO:0016705	0-3, 3-3.5, 3.5-4.5, 4.5-5.5	MF	oxidoreductase activity, acting on paired donors, with
GO:0016747	0-3, 3-3.5, 3.5-4.5, 4.5-5.5	MF	transferase activity, transferring acyl groups other than
GO:0016301	0-3, 3-3.5, 4.5-5.5	MF	kinase activity
GO:0022891	0-3, 3-3.5, 4.5-5.5	MF	substrate-specific transmembrane transporter activity
GO:0003700	0-3, 3-3.5,3.5-4.5,4.5-5.5	MF	sequence-specific DNA binding transcription factor activity
GO:0004497	0-3, 3-3.5,3.5-4.5,4.5-5.5	MF	monooxygenase activity
GO:0016760	0-3, 3.5-4.5	MF	cellulose synthase (UDP-forming) activity
GO:0003677	3-3.5	MF	DNA binding
GO:0003838	3-3.5	MF	sterol 24-C-methyltransferase activity
GO:0003916	3-3.5	MF	DNA topoisomerase activity
GO:0003918	3-3.5	MF	DNA topoisomerase (ATP-hydrolyzing) activity
GO:0004190	3-3.5	MF	aspartic-type endopeptidase activity
GO:0004356	3-3.5	MF	glutamate-ammonia ligase activity
GO:0004503	3-3.5	MF	monophenol

			monooxygenase activity
GO:0004714	3-3.5	MF	transmembrane receptor protein tyrosine kinase activity
GO:0004857	3-3.5	MF	enzyme inhibitor activity
GO:0005199	3-3.5	MF	structural constituent of cell wall
GO:0005200	3-3.5	MF	structural constituent of cytoskeleton
GO:0005351	3-3.5	MF	sugar
GO:0008378	3-3.5	MF	galactosyltransferase activity
GO:0008509	3-3.5	MF	anion transmembrane transporter activity
GO:0008569	3-3.5	MF	minus-end-directed microtubule motor activity
GO:0009678	3-3.5	MF	hydrogen-translocating pyrophosphatase activity
GO:0015035	3-3.5	MF	protein disulfide oxidoreductase activity
GO:0015144	3-3.5	MF	carbohydrate transmembrane transporter activity
GO:0016157	3-3.5	MF	sucrose synthase activity
GO:0016746	3-3.5	MF	transferase activity, transferring acyl groups
GO:0016818	3-3.5	MF	hydrolase activity, acting on acid anhydrides, in
GO:0016866	3-3.5	MF	intramolecular transferase activity
GO:0019899	3-3.5	MF	enzyme binding
GO:0030247	3-3.5	MF	polysaccharide binding
GO:0030674	3-3.5	MF	protein binding, bridging
GO:0045735	3-3.5	MF	nutrient reservoir activity

Table B.1 Continued

GO:0046982	3-3.5	MF	protein heterodimerization activity
GO:0047262	3-3.5	MF	polygalacturonate 4-alpha-galacturonosyltransferase activity
GO:0047672	3-3.5	MF	anthranilate N-benzoyltransferase activity
GO:0051015	3-3.5	MF	actin filament binding
GO:0070566	3-3.5	MF	adenylyltransferase activity
GO:0080116	3-3.5	MF	glucuronoxylan glucuronosyltransferase activity
GO:0080123	3-3.5	MF	jasmonate-amino synthetase activity
GO:0003824	3-3.5, 3.5-4.5	MF	catalytic activity
GO:0003886	3-3.5, 3.5-4.5	MF	DNA (cytosine-5-)-methyltransferase activity
GO:0004353	3-3.5, 3.5-4.5	MF	glutamate dehydrogenase [NAD(P)+] activity
GO:0004478	3-3.5, 3.5-4.5	MF	methionine adenosyltransferase activity
GO:0004713	3-3.5, 3.5-4.5	MF	protein tyrosine kinase activity
GO:0005507	3-3.5, 3.5-4.5	MF	copper ion binding
GO:0008017	3-3.5, 3.5-4.5	MF	microtubule binding
GO:0008171	3-3.5, 3.5-4.5	MF	O-methyltransferase activity
GO:0008422	3-3.5, 3.5-4.5	MF	beta-glucosidase activity
GO:0008474	3-3.5, 3.5-4.5	MF	palmitoyl-(protein) hydrolase activity
GO:0008725	3-3.5, 3.5-4.5	MF	DNA-3-methyladenine glycosylase activity

Table B.1 Continued

GO:0008810	3-3.5, 3.5-4.5	MF	cellulase activity
GO:0016639	3-3.5, 3.5-4.5	MF	oxidoreductase activity, acting on the CH-NH2 group of donors,
GO:0016787	3-3.5, 3.5-4.5	MF	hydrolase activity
GO:0016841	3-3.5, 3.5-4.5	MF	ammonia-lyase activity
GO:0019901	3-3.5, 3.5-4.5	MF	protein kinase binding
GO:0042349	3-3.5, 3.5-4.5	MF	guiding stereospecific synthesis activity
GO:0045548	3-3.5, 3.5-4.5	MF	phenylalanine ammonia-lyase activity
GO:0003777	3-3.5, 3.5-4.5	MF	microtubule motor activity
GO:0004097	3-3.5, 3.5-4.5-4.5-5.5	MF	catechol oxidase activity
GO:0004601	3-3.5, 3.5-4.5, 4.5-5.5	MF	peroxidase activity
GO:0005315	3-3.5, 3.5-4.5, 4.5-5.5	MF	inorganic phosphate transmembrane transporter activity
GO:0008974	3-3.5, 3.5-4.5, 4.5-5.5	MF	phosphoribulokinase activity
GO:0010333	3-3.5, 3.5-4.5, 4.5-5.5	MF	terpene synthase activity
GO:0016165	3-3.5, 3.5-4.5, 4.5-5.5	MF	lipoxygenase activity
GO:0016168	3-3.5, 3.5-4.5, 4.5-5.5	MF	chlorophyll binding
GO:0016298	3-3.5, 3.5-4.5, 4.5-5.5	MF	lipase activity
GO:0016757	3-3.5, 3.5-4.5, 4.5-5.5	MF	transferase activity, transferring glycosyl groups
GO:0016758	3-3.5, 3.5-4.5, 4.5-5.5	MF	transferase activity, transferring hexosyl groups
GO:0016762	3-3.5, 3.5-4.5, 4.5-5.5	MF	xyloglucan
GO:0016788	3-3.5, 3.5-4.5, 4.5-5.5	MF	hydrolase activity, acting on ester bonds

Table B.1 Continued

GO:0016798	3-3.5, 3.5-4.5, 4.5-5.5	MF	hydrolase activity, acting on glycosyl bonds
GO:0020037	3-3.5, 3.5-4.5, 4.5-5.5	MF	heme binding
GO:0030246	3-3.5, 3.5-4.5, 4.5-5.5	MF	carbohydrate binding
GO:0030599	3-3.5, 3.5-4.5, 4.5-5.5	MF	pectinesterase activity
GO:0045330	3-3.5, 3.5-4.5, 4.5-5.5	MF	aspartyl esterase activity
GO:0046983	3-3.5, 3.5-4.5, 4.5-5.5	MF	protein dimerization activity
GO:0052716	3-3.5, 3.5-4.5, 4.5-5.5	MF	hydroquinone
GO:0022857	3-3.5, 4.5-5.5	MF	transmembrane transporter activity
GO:0050403	3-3.5, 4.5-5.5	MF	trans-zeatin O-beta-D-glucosyltransferase activity
GO:0050664	3-3.5, 4.5-5.5	MF	oxidoreductase activity, acting on NADH or NADPH, oxygen as
GO:0003885	3.5-4.5	MF	D-arabinono-1,4-lactone oxidase activity
GO:0004180	3.5-4.5	MF	carboxypeptidase activity
GO:0004190	3.5-4.5	MF	aspartic-type endopeptidase activity
GO:0004351	3.5-4.5	MF	glutamate decarboxylase activity
GO:0004352	3.5-4.5	MF	glutamate dehydrogenase (NAD+) activity
GO:0004365	3.5-4.5	MF	glyceraldehyde-3-phosphate dehydrogenase (NAD+) (phosphorylating)
GO:0004435	3.5-4.5	MF	phosphatidylinositol phospholipase C activity
GO:0004470	3.5-4.5	MF	malic enzyme activity

Table B.1 Continuted

GO:0004664	3.5-4.5	MF	prephenate dehydratase activity
GO:0004857	3.5-4.5	MF	enzyme inhibitor activity
GO:0005200	3.5-4.5	MF	structural constituent of cytoskeleton
GO:0005388	3.5-4.5	MF	calcium-transporting ATPase activity
GO:0008081	3.5-4.5	MF	phosphoric diester hydrolase activity
GO:0008661	3.5-4.5	MF	1-deoxy-D-xylulose-5-phosphate synthase activity
GO:0008762	3.5-4.5	MF	UDP-N-acetylmuramate dehydrogenase activity
GO:0015385	3.5-4.5	MF	sodium
GO:0016210	3.5-4.5	MF	naringenin-chalcone synthase activity
GO:0016614	3.5-4.5	MF	oxidoreductase activity, acting on CH-OH group of donors
GO:0016619	3.5-4.5	MF	malate dehydrogenase (oxaloacetate-decarboxylating) activity
GO:0016620	3.5-4.5	MF	oxidoreductase activity, acting on the aldehyde or oxo group of
GO:0032440	3.5-4.5	MF	2-alkenal reductase [NAD(P)] activity
GO:0033843	3.5-4.5	MF	xyloglucan 6-xylosyltransferase activity
GO:0046577	3.5-4.5	MF	long-chain-alcohol oxidase activity
GO:0050660	3.5-4.5	MF	flavin adenine dinucleotide binding
GO:0050661	3.5-4.5	MF	NADP binding
GO:0051119	3.5-4.5	MF	sugar transmembrane transporter activity
GO:0051287	3.5-4.5	MF	NAD binding

Table B.1 Continued

GO:0004611	3.5-4.5, 4.5-5.5	MF	phosphoenolpyruvate carboxykinase activity
GO:0010279	3.5-4.5, 4.5-5.5	MF	indole-3-acetic acid amido synthetase activity
GO:0016702	3.5-4.5, 4.5-5.5	MF	oxidoreductase activity, acting on single donors with
GO:0016829	3.5-4.5, 4.5-5.5	MF	lyase activity
GO:0016831	3.5-4.5, 4.5-5.5	MF	carboxy-lyase activity
GO:0016851	3.5-4.5, 4.5-5.5	MF	magnesium chelatase activity
GO:0017076	3.5-4.5, 4.5-5.5	MF	purine nucleotide binding
GO:0035252	3.5-4.5, 4.5-5.5	MF	UDP-xylosyltransferase activity
GO:0043531	3.5-4.5, 4.5-5.5	MF	ADP binding
GO:0043565	3.5-4.5, 4.5-5.5	MF	sequence-specific DNA binding
GO:0046863	3.5-4.5, 4.5-5.5	MF	ribulose-1,5-bisphosphate carboxylase/oxygenase activator
GO:0050242	3.5-4.5, 4.5-5.5	MF	pyruvate, phosphate dikinase activity
GO:0000822	4.5-5.5	MF	inositol hexakisphosphate binding
GO:0003676	4.5-5.5	MF	nucleic acid binding
GO:0003964	4.5-5.5	MF	RNA-directed DNA polymerase activity
GO:0004190	4.5-5.5	MF	aspartic-type endopeptidase activity
GO:0004367	4.5-5.5	MF	glycerol-3-phosphate dehydrogenase [NAD+] activity
GO:0004451	4.5-5.5	MF	isocitrate lyase activity
GO:0004519	4.5-5.5	MF	endonuclease activity
GO:0004523	4.5-5.5	MF	ribonuclease H activity

Table B.1 Continued

GO:0004612	4.5-5.5	MF	phosphoenolpyruvate carboxykinase (ATP) activity
GO:0004857	4.5-5.5	MF	enzyme inhibitor activity
GO:0008137	4.5-5.5	MF	NADH dehydrogenase (ubiquinone) activity
GO:0008270	4.5-5.5	MF	zinc ion binding
GO:0008271	4.5-5.5	MF	secondary active sulfate transmembrane transporter activity
GO:0008519	4.5-5.5	MF	ammonium transmembrane transporter activity
GO:0010181	4.5-5.5	MF	FMN binding
GO:0010329	4.5-5.5	MF	auxin efflux transmembrane transporter activity
GO:0015020	4.5-5.5	MF	glucuronosyltransferase activity
GO:0015116	4.5-5.5	MF	sulfate transmembrane transporter activity
GO:0015299	4.5-5.5	MF	solute
GO:0016040	4.5-5.5	MF	glutamate synthase (NADH) activity
GO:0016630	4.5-5.5	MF	protochlorophyllide reductase activity
GO:0016887	4.5-5.5	MF	ATPase activity
GO:0019825	4.5-5.5	MF	oxygen binding
GO:0033897	4.5-5.5	MF	ribonuclease T2 activity
GO:0042132	4.5-5.5	MF	fructose 1,6-bisphosphate 1-phosphatase activity
GO:0042578	4.5-5.5	MF	phosphoric ester hydrolase activity
GO:0045181	4.5-5.5	MF	glutamate synthase activity, NADH or NADPH as acceptor

Table B.1 Continued

GO:0045550	4.5-5.5	MF	geranylgeranyl reductase activity
GO:0046857	4.5-5.5	MF	oxidoreductase activity, acting on other nitrogenous compounds as
GO:0046872	4.5-5.5	MF	metal ion binding
GO:0047750	4.5-5.5	MF	cholestenol delta-isomerase activity
GO:0047787	4.5-5.5	MF	delta4-3-oxosteroid 5beta-reductase activity
GO:0000220	3-3.5	CC	vacuolar proton-transporting V-type ATPase, V0 domain
GO:0000228	4.5-5.5	CC	nuclear chromosome
GO:0000325	3-3.5	CC	plant-type vacuole
GO:0000786	3-3.5	CC	nucleosome
GO:0000786	3.5-4.5	CC	nucleosome
GO:0000796	3-3.5	CC	condensin complex
GO:0005576	3-3.5	CC	extracellular region
GO:0005576	3.5-4.5	CC	extracellular region
GO:0005576	4.5-5.5	CC	extracellular region
GO:0005618	3-3.5	CC	cell wall
GO:0005618	3.5-4.5	CC	cell wall
GO:0005819	3.5-4.5	CC	spindle
GO:0005871	3-3.5	CC	kinesin complex
GO:0005871	3.5-4.5	CC	kinesin complex
GO:0005874	3-3.5	CC	microtubule
GO:0005874	3.5-4.5	CC	microtubule
GO:0005875	3-3.5	CC	microtubule associated complex
GO:0005875	3.5-4.5	CC	microtubule associated complex
GO:0005887	3.5-4.5	CC	integral to plasma membrane
GO:0005971	0-3	CC	ribonucleoside-diphosphate reductase complex
GO:0009331	4.5-5.5	CC	glycerol-3-phosphate dehydrogenase complex
GO:0009505	3-3.5	CC	plant-type cell wall

Table B.1 Continued

GO:0009505	3.5-4.5	CC	plant-type cell wall
GO:0009505	4.5-5.5	CC	plant-type cell wall
GO:0009522	3-3.5	CC	photosystem I
GO:0009522	3.5-4.5	CC	photosystem I
GO:0009522	4.5-5.5	CC	photosystem I
GO:0009523	3-3.5	CC	photosystem II
GO:0009523	3.5-4.5	CC	photosystem II
GO:0009523	4.5-5.5	CC	photosystem II
GO:0009524	3-3.5	CC	phragmoplast
GO:0009524	3.5-4.5	CC	phragmoplast
GO:0009535	3.5-4.5	CC	chloroplast thylakoid membrane
GO:0009536	4.5-5.5	CC	plastid
GO:0009538	3-3.5	CC	photosystem I reaction center
GO:0009538	3.5-4.5	CC	photosystem I reaction center
GO:0009538	4.5-5.5	CC	photosystem I reaction center
GO:0009543	3.5-4.5	CC	chloroplast thylakoid lumen
GO:0009579	3.5-4.5	CC	thylakoid
GO:0009579	4.5-5.5	CC	thylakoid
GO:0009654	3.5-4.5	CC	oxygen evolving complex
GO:0009654	4.5-5.5	CC	oxygen evolving complex
GO:0009705	3-3.5	CC	plant-type vacuole membrane
GO:0009705	4.5-5.5	CC	plant-type vacuole membrane
GO:0015935	0-3	CC	small ribosomal subunit
GO:0016020	3-3.5	CC	membrane
GO:0016020	3.5-4.5	CC	membrane
GO:0016020	4.5-5.5	CC	membrane
GO:0016021	3-3.5	CC	integral to membrane
GO:0016021	3.5-4.5	CC	integral to membrane
GO:0016021	4.5-5.5	CC	integral to membrane
GO:0016023	3.5-4.5	CC	cytoplasmic membrane-bounded vesicle

Table B.1 Continued

GO:0016459	4.5-5.5	CC	myosin complex
GO:0030095	3.5-4.5	CC	chloroplast photosystem II
GO:0031225	3-3.5	CC	anchored to membrane
GO:0031225	3.5-4.5	CC	anchored to membrane
GO:0031977	0-3	CC	thylakoid lumen
GO:0031977	3.5-4.5	CC	thylakoid lumen
GO:0042555	0-3	CC	MCM complex
GO:0043234	3.5-4.5	CC	protein complex
GO:0045263	4.5-5.5	CC	proton-transporting ATP synthase complex, coupling factor F(o)
GO:0046658	0-3	CC	anchored to plasma membrane
GO:0046658	3-3.5	CC	anchored to plasma membrane
GO:0046658	3.5-4.5	CC	anchored to plasma membrane
GO:0046658	4.5-5.5	CC	anchored to plasma membrane
GO:0048046	3-3.5	CC	apoplast
GO:0048046	3.5-4.5	CC	apoplast
GO:0048046	4.5-5.5	CC	apoplast
GO:0070469	4.5-5.5	CC	respiratory chain

VITA

VITA

Nadia Marie Atallah

EDUCATION

2011-Present Ph.D., Botany & Plant Pathology, Purdue University

- Doctoral Degree expected in May 2015
- Advisor: Dr. Jo Ann Banks, Purdue University, Botany and Plant Pathology
- GPA: 3.90
- Certification in Computational Life Sciences (CLS)
    - Related Coursework: Statistical Methods for Bioinformatics and Computational Biology, Applied Linear Regression Analysis, Statistical Methods for Biology, Practical Biocomputing, Experimental Design, Introduction to Algorithms

2011 B.A. in Psychology, Minor: Law and Society, Purdue University

2011 B.S. in Biochemistry, Purdue University

RESEARCH EXPERIENCE

10/2011–Present - Research Assistant, Department of Botany & Plant Pathology, Purdue University. Mentor: Dr. Jo Ann Banks: Investigating sex determination and differentiation in the homosporous fern *Ceratopteris richardii*. Comparing male and hermaphrodite Ceratopteris gametophytes using RNA-Seq, including *de novo* assembly of the transcriptome, quantification of sequence reads, differential expression analysis with R & Bioconductor packages, identification of differentially expressed genes, and using RNAi to knock down candidate and differentially expressed genes. This study provides the first comprehensive transcriptome of *Ceratopteris* and the first genomics approach to determine differential gene expression of male and hermaphrodite gametophtyes. The transcriptome is already being used by several research groups in phylogenetics. The differential expression analysis showed that sex determination is regulated by multiple processes and involves the interplay of phytohormones, protein turnover, and epigenetic remodeling of the genome. An additional time-course has been performed to observe gene expression profiles of germinating spores across early development. These RNA-Seq experiments have enabled ongoing experiments with gene knock-downs using RNAi. Ceratopteris is a powerful model system for studying sex

determination in land plants. We know exactly when, where, and how sex is determined. Numerous sex determining mutants have been identified in Ceratopteris and through test of epistasis, a genetic sex determination pathway has been described. Ceratopteris is an ideal organism for studying sex determination and these experiments pave the way to obtaining a greater understanding of sex determination in plants.

5/20/2013-8/16/2013 - Intern, Dow AgroSciences, Discovery Plant Pathology. Research Scientist: Dr. Javier Delgado: Developed a SOP and a user guide for new metabolic profiling analytical machinery and software; trained DAS scientists to use the system. Optimized the system for mode-of-action determination, including compound concentration selection, pathogen selection, software settings, and statistical analysis workflow used. Used critical thinking and troubleshooting to identify tools and workflows for analyzing data. Presented findings in both a poster presentation and group meeting. Used computational biology skills to investigate the utility of RNAi for control of plant pathogens. Identified RNAi machinery in multiple fungal genomes.

01/2010 – 08/2010 Undergraduate Researcher, Department of Biochemistry, Purdue University. Mentor: Dr. Elizabeth Tran: Investigated the effects of DEAD-box protein Dbp2 point-mutations in *Saccharomyces cerevisiae*. Developed and isolated novel dominant negative mutants using hydroxylamine mutagenesis. Biochemical methods used included molecular cloning, lithium acetate yeast transformation, and western blotting.

AWARDS & HONORS

2014	Botany & Plant Pathology Travel Grant
2013-Present	Phi Kappa Phi Honor Society
2012	Botany & Plant Pathology Travel Grant
2009-2011	Dean's List of Distinguished Students

PROFESSIONAL SERVICE

2014-present *Ad hoc* reviewer for *Plant Cell*

LEADERSHIP, TEACHING & MENTORING

Teaching Assistantships

Fall 2011	BTNY 305, Plant Systematics Teaching assistant, Purdue University
Fall 2012	BTNY 305, Plant Systematics Teaching assistant, Purdue University
Fall 2013	BTNY 305, Plant Systematics Teaching assistant, Purdue University
Fall 2013	STAT598C bioinformatics guest lecture on RNA-Seq experimental design and data analysis, invited by Dr. Olga Vitek

- In addition to teaching assistant responsibilities, lectured on phylogenetics and basic informatics tools (such as BLAST).
- I developed a prairie plant identification field trip for our students: contacted Niches landtrust, planned and organized the field trip, and drove and monitored students during the trip**.**
- I participated in the organization of the 2014 Purdue University Botany & Plant Pathology/Entomology trip to Dow AgroSciences headquarters. I disseminated information at Purdue, organized participants, organized and secured travel to Indianapolis.

Banks lab students mentored:
Mentored students in data analysis and laboratory techniques

- JaLeah Hendricks, Katherine Embry, Kye Stachowski, Andrew Eller, Stephan Mielke, Barbara Dale, Yuchen Gang

Data analysis aid/bioinformatics mentoring:
Aided several research groups in RNA-Seq analysis and workflow design, fixed R scripts, and wrote custom Perl scripts.

- Siwen Wang, graduate student, advisor: Dr. Elizabeth Tran, Purdue University
- Gabriel Patrick Hughes, graduate student, advisor: Dr. Matthew Ginzel, Purdue University
- Micah Stevens, graduate student, advisor: Dr. Paula Pijut, Purdue University
- Guotian Li, graduate student, advisor: Dr. Jin-Rong Xu, Purdue University
- Dr. Scott McAdams, post doctoral researcher, post doctoral advisor: Dr. Tim Brodribb
- Archana Chauhan, post doctoral researcher, Center for Environmental Biotechnology, University of Tennessee

BIOINFORMATIC, COMPUTER, & LABORATORY EXPERTISE

Laboratory:
RNA extraction, tissue culture (*Ceratopteris*, yeast, qRT-PCR, DNA extraction, restriction enzyme digests, Gateway cloning, mutagenesis, yeast transformation, bacterial transformation, media preparation, microscopy, plasmid extraction, competent cell preparation, cDNA synthesis, 5'-RACE, Northern blot, PCR, Biolistic bombardment, RNAi, Agrobacterium-mediated transformation, Hairpin-vector construction, Molecular cloning, Western blotting, Biolog metabolic profiling, gel extraction, sterile technique

Computer:
Mac OS, Windows and MS Office suite, UNIX/LINUX
Statistical packages: SAS, JMP
Programming languages: Perl, R, HTML, Bash UNIX shell script

Bioinformatics Software:
vsn, edgeR, DESeq, DESeq2, EBSeq, DeconSeq, MEGA, BLAST, BLAST2GO, Trinity, RSEM, Bowtie, Trimmomatic, FastQC, affy, limma, MrBayes, genefilter, ggplot2, PFAM, Cytoscape, Inkscape, GOSeq, FASTX, BWA, AgriGO, ClustalX, TreeGraph 2, DAVID

Bioinformatics Analyses:
RNA-Seq data analysis (with and without a reference genome), microarray analysis, experimental design, unsupervised data exploration, pathway analysis, phylogenetics, sequence annotation, GO enrichment analysis

PUBLICATIONS

Invited Reviews:

- **Atallah NM** and Banks J (2015). Reproduction and the pheromonal regulation of sex type in fern gametophytes..*Front. Plant Sci.* **6**:100. doi: 10.3389/fpls.2015.00100

MANUSCRIPTS IN PREPARATION

Research papers:

- **Atallah NM**, Gribskov M, Vitek O, Gaiti F, Banks JA, Tanurdzic M. (2014). Transcriptional reprogramming of *Ceratopteris richardii* gametophytes by the sex determining pheromone antheridiogen.
- Wu Q, DeLeon A, **Atallah NM**, Gribskov M, Banks JA. (2014). Transcriptome assembly and differential expression analysis of *Pteris vittata* in response to arsenic.
- Hass B, **Atallah NM**, Banks JA. (2014). Insight into the function of ABA in Ceratopteris gametophytes through ABA mutant analysis.
- **Atallah NM**, Vitek O, Banks JA. (2015) Temporal gene expression profile in developing *Ceratopteris richardii* gametophytes.

PRESENTATIONS

1. Atallah N, Tanurdzic M, Gribskov M, Vitek O, Banks JA. Sex Determination and Transcriptional Reprogramming of *Ceratopteris richardii* Gametophytes by a GA-like Pheromone. Poster presented at: Botany and Plant Pathology Poster Session; 2014; West Lafayette, IN.
2. Purdue Zip Trips, 2014, “It’s a Gene Thing”,
3. https://www.youtube.com/watch?v=9x_MzWJYbnk&feature=share
4. Atallah N, Tanurdzic M, Gribskov M, Vitek O, Banks JA. Identifying genes involved in sex determination in *Ceratopteris richardii*. Poster presented at: Botany and Plant Pathology Poster Session; 2013; West Lafayette, IN.
5. Atallah N, Tanurdzic M, Gribskov M, Vitek O, Banks JA. Identifying genes involved in sex determination in *Ceratopteris richardii.* Poster presented at: Botany and Plant Pathology Poster Session; 2012; West Lafayette, IN.

6. Atallah N, Tanurdzic M, Gribskov M, Vitek O, Banks JA. Identifying genes involved in sex determination in *Ceratopteris richardii.* Oral presentation at: Botany and Plant Pathology Poster Session; 2012; West Lafayette, IN.
7. Atallah N, Tanurdzic M, Vitek O, Banks JA. Identifying sex determination genes in the fern *Ceratopteris richardii*. Poster presented at: 77th Symposium at Cold Spring Harbob Laboratory: The Biology of Plants; 2012 May30-Jun4; Cold Spring Harbor, NY.
8. Atallah N, Tanurdzic M, Gribskov M, Vitek O, Banks JA. Differential Gene Expression in Fern Gametophytes Using RNAseq. Invited speaker at: Midwest Illumina User Group; 2012, Oct. 24,25; St. Louis, MO.
9. Atallah N, Tanurdzic M, Vitek O, Banks JA. Identifying genes involved in sex determination in *Ceratopteris richardii*. Poster presented at: Botany and Plant Pathology Poster Session; 2011; West Lafayette, IN.

PROFESSIONAL MEETINGS AND CONFERENCES ATTENDED

Plant Biology:

- ASPB 2011 conference (Minneapolis, Minnesota);, 8/6/2011-8/10/2011
- 77th Symposium: The Biology of Plants (Cold Spring Harbor Laboratories, NY), 5/30/2012-6/4/2012
    - Attended, volunteered, and presented poster.

Bioinformatics:

- Plant Genomes & Biotechnology conference (Cold Spring Harbor Laboratories, NY), 11/30/2011-12/3/2011
    - Attended and participated in iPlant Collaborative workshop on high-throughput data analysis.
- Midwest Illumina User Group (St. Louis, MO), 10/24/2012 & 10/25/2012;
    - Invited speaker, Title: "Differential Expression in Fern Gametophytes using RNA-Seq"
- Great Lakes Bioinformatics Conference (Cincinnati, OH), May16-18, 2014;
    - Attended and participated in workshops; the majority of this conference was oncology-related

BIOINFORMATIC WORKSHOPS ATTENDED

2011 iPlant Collaborative Workshop on high-throughput data analysis, Cold Spring Harbor Laboratories, NY

2012 Microarray Data Analysis online workshop through Canadian Bioinformatics Workshops

2014 Lincs Chemical Biology Data Analysis Workshop, Cincinnati Children's Hospital Medical Center, Cincinnati, OH

2014 CanvasXpress: PCalign: A Method to Quantify Physiochemical Similarity of Protein-Protein Interfaces, Cincinnati Children's Hospital Medical Center, Cincinnati, OH

2014 Enabling Collaborative Research Through Synapse: A Cloud Environment for Data Sharing and Analysis, Cincinnati Children's Hospital Medical Center, Cincinnati, OH

2014 Methods and Approaches for the Analysis of Gene Signaling Pathways and Disease Gene Ranking, Cincinnati Children's Hospital Medical Center, Cincinnati, OH

REFERENCES

Dr. Jody Banks
Department of Botany and Plant Pathology
Purdue University
915 West State St., West Lafayette, IN 47907
Email: banksj@purdue.edu

Dr. Javier Delgado
Discovery, Plant Pathology
Dow AgroSciences LLC
9330 Zionsville Road
Indianapolis, IN 46268
Email: JADelgado@dow.com

Dr. Olga Vitek
College of Information and Computer Science
Northeastern University
202 West Village H, Boston, MA 02115
Email: ovitek@stat.purdue.edu

Dr. Michael Gribskov
Department of Biological Sciences & Department of Computer Science
Purdue University
915 West State St., West Lafayette, IN 47907
Email: grisbskov@purdue.edu